NASA CR-
147763

(NASA-CR-147763) NONPARAMETRIC PROBABILITY DENSITY ESTIMATION BY OPTIMIZATION THEORETIC TECHNIQUES (RICE UNIV.) 156 P HC $6.75 CSCL 12A

N76-25890

UNCLAS
G3/65 42587

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Nonparametric Probability Density Estimation by Optimization Theoretic Techniques

by

David Warren Scott
Mathematical Sciences Department
Rice University

ABSTRACT

In this study two nonparametric probability density estimators are considered. The first is the kernel estimator. The problem of choosing the kernel scaling factor based solely on a random sample is addressed. An interactive mode is discussed and an algorithm proposed to choose the scaling factor automatically. In a Monte Carlo simulation study, the resulting integrated mean square error compares favorably with the error using the usual asymptotically optimal choice of the kernel scaling factor. For the latter case, the true sampling density is required to calculate the optimal scaling factor.

The second nonparametric probability estimate uses penalty function techniques with the maximum likelihood criterion. A discrete maximum penalized likelihood estimator is proposed and is shown to be consistent in the mean square error. Approximation results of this discrete solution to the corresponding infinite-dimensional solution are proved. A numerical implementation technique for the discrete solution is discussed and examples displayed. An extensive simulation study compares the integrated mean square error of the discrete and kernel estimators. The robustness of the discrete estimator is demonstrated graphically.

Institute for Computer Services and Applications
Rice University
Houston, Texas
April, 1976

Research supported under Office of Naval Research Grant N00014-75-C-0452
and under NASA Contract NAS 9-12776

ACKNOWLEDGMENTS

I would like to thank my thesis advisors, Professors James R. Thompson and Richard A. Tapia for their guidance and inspiration. I would also like to acknowledge many helpful discussions with my other two committee members, Professors Paul E. Pfeiffer and David A. Schum. The discussions with my fellow students, Messrs. R. Byrd, L. Zyla, and R. Grisham were most helpful.

To my wife, Jean, I express my deepest appreciation for her encouragement and patience. I also wish to thank my parents for their confidence and support throughout my graduate years.

Finally, for her careful typing, I am grateful to Ms. Louise Power. Part of this work was supported under Grant #N00014-75-C-0452 from the Office of Naval Research. Computer funds were made available by NASA under Grant NAS9-12776, for which I acknowledge Dr. M.S. Lynn and Dr. D. Van Rooy.

TABLE OF CONTENTS

I. INTRODUCTION: THE PROBLEM OF PROBABILITY DENSITY ESTIMATION

The probabilistic nature of our world is a feature which mankind has had to cope with throughout his existence. Only comparatively recently has he attempted to cope with the uncertainty in a formal fashion. In many situations it is desirable to study the underlying stochastic structure by specifying a probability density function which reflects the random behavior of the data. Thus the problem of estimating a probability density function from a set of data is of extreme importance.

Specifically, we wish to find a function $\hat{f}(\cdot)$ which is an estimate for an unknown probability density function $f(\cdot)$, based on a random sample $x_1, x_2, \ldots, x_N$ from $f(\cdot)$. The methodologies for solving this problem fall into two general classes, parametric and nonparametric procedures. Actually, this division in practice is not sharp. In a sense there is a series of steps from an assumption of a specific functional form of the probability density $f(\cdot)$ to much weaker assumptions, for example $f \in C^2$ with $\int f''(x)^2 dx \leq k$. We should be aware of the fallacy of the belief that any nonparametric procedure is "assumption free."

In the following sections we examine the available methods and a new nonparametric algorithm for estimating probability densities. Computer examples are presented and Monte Carlo simulations summarized to evaluate the performances of several algorithms. Particular attention is paid to the theoretical numerical analytical and statistical properties of the new algorithm. We begin by examining the philosophy of commonly used techniques for the two general approaches for estimating probability density functions.

1.1 Parametric Estimation

For parametric estimation it is assumed that the unknown sampling density takes a known functional form

$$f(x) \equiv f(x|\theta)$$

where θ is a vector of p parameters which completely specifies the density function. Given such a functional representation of the density, the parametric form of the sampling density is known. Thus the problem of estimating the density function involves the estimation of components of the vector θ.

In restricting the class of possible densities in this parametric fashion, it is clear that our estimates may not be robust against an incorrect assumption of the parametric class. For example, we might assume a unimodal density while the true density is bimodal. Clearly, prior knowledge of the density's explicit functional form is extremely useful.

There are several popular parametric estimation procedures for choosing statistics to estimate the unknown parameter θ. Frequently, for example, a statistic Y is sought that is unbiased and has minimum mean square error subject to this constraint.

A Bayesian Procedure

Let us consider a Bayesian method for injecting prior knowledge of θ into the estimation procedure. For simplicity, let θ be a single parameter which takes on values in the interval (a,b). We suppose that the knowledge of the true value of θ is characterized by a prior probability density $\lambda(\theta)$. The joint density of θ and a random sample $x = \{x_1,\ldots,x_N\}$ is given by

$$h(x,\theta) = \prod_{i=1}^{N} f(x_i|\theta)\lambda(\theta) = L(\theta|x)\lambda(\theta)$$

where $L(\theta|x)$ is called the likelihood function. Applying Bayes Theorem, we may calculate the conditional density of θ given x or the posterior density of θ as

$$g(\theta|x) = \frac{L(\theta|x)\lambda(\theta)}{\int L(\theta'|x)\lambda(\theta')d\theta'} .$$

There are several possible estimates $\hat{\theta}(x)$ for θ based on the posterior density $g(\theta|x)$. We may use the median or mode of $g(\theta|x)$ to estimate the parameter θ, although the mode may not be unique. Alternately, we may consider the estimator $\hat{\theta}(x)$ which minimizes the criterion function

$$E[(\hat{\theta}(x)-\theta)^2] = \int_x \int_\theta (\hat{\theta}(x)-\theta)^2 \lambda(\theta)L(\theta|x)d\theta dx .$$

We may do so by assigning to $\hat{\theta}(x)$ for each x that value which minimizes

$$\int_\theta (\hat{\theta}(x)-\theta)^2 L(\theta|x)\lambda(\theta)d\theta .$$

Differentiating with respect to the value $\hat{\theta}(x)$ and setting the derivative equal to zero implies

$$\hat{\theta}(x) = \int \theta g(\theta|x)d\theta$$

that is, $\hat{\theta}(x)$ is simply the mean of the posterior density $g(\theta|x)$.

A Maximum Likelihood Procedure

In 1922, R.A. Fisher [1950] introduced a new criterion for choosing the parameter θ which he called the maximum likelihood estimate. Here, θ is chosen to maximize the likelihood function $L(\theta|x)$. We motivate Fisher's estimator with a Bayesian argument. Suppose we restrict θ to a finite interval (a,b) and consider a prior density for θ that is constant on (a,b),

$$\lambda(\theta) = \begin{cases} \frac{1}{b-a} & \text{if } a < \theta < b \\ 0 & \text{otherwise} \end{cases} .$$

Hence we assume that every point in (a,b) is equally likely to be the true value of θ according to our prior knowledge. Thus the posterior density for $\theta \in (a,b)$ is

$$g(\theta|x) = \frac{L(\theta|x)}{\int_a^b L(\theta'|x)d\theta'} .$$

If we consider the posterior mode of $g(\theta|x)$, we see, since the denominator does not depend on θ after performing the integration, that maximizing $g(\theta|x)$ is equivalent to maximizing the likelihood function $L(\theta|x)$. Thus the maximum likelihood estimate, according to the Bayesian interpretation, is the mode of the posterior density $g(\theta|x)$ assuming a uniform prior density on θ.

Maximum likelihood estimates will not generally be unique, but under certain regularity assumptions about $f(.|\theta)$, Huzurbazar [1948] and Wald [1949] have shown that the maximum likelihood estimators are unique and consistent as the sample size N tends to infinity. The Bayesian interpretation of the maximum likelihood estimate was rejected by Fisher. The use of the maximum likelihood philosophy in the nonparametric setting has been the subject of much recent work, including this thesis.

We summarize our discussion of parametric probability density function estimates by emphasizing their importance in modelling and their relative efficiency under the correct hypotheses. However, we warn that these procedures are not robust against errors in choosing the parametric family.

1.2 Nonparametric Estimation

In 1895 Karl Pearson [1948] proposed a systematic method for fitting a probability density function based on the first four sample moments of a random sample. Motivated by a limiting form of the hypergeometric distribution, he proposed choosing the density $f(x)$ which solves the differential equation

$$\frac{d \log f(x)}{dx} = \frac{x - a}{b_o + b_1 x + b_2 x^2} \tag{1.2.1}$$

where a, b_o, b_1, and b_2 depend on the first four sample moments and $b_1 = -a$. The Normal, Gamma, Beta, F, and Student's t distributions are members of the Pearsonian family of densities. Unfortunately, the differential equation has three independent parameters. However, most of the univariate densities mentioned above have no more than two determining parameters. Thus the probability is zero that a solution to (1.2.1) will be, say, a normal density.

For our purposes, we define a nonparametric probability density estimator as one that does not result from an a priori choice of the parametric form of a known density. The advantage of a nonparametric estimator is that it can approximate a wide range of true densities, whereas we are committed under the assumption of a parametric density form.

Pearson's estimation procedure is, by our definition, clearly parametric. However, it admits of a more general class of density estimates than if we assumed a priori that, say, the unknown density is Gaussian with unknown mean and variance. Pearson's family of density estimates is itself reasonably restrictive. For example, it contains no densities with more than one internal mode.

We shall consider other nonparametric estimators, beginning with the

histogram estimator. The kernel estimator will be considered in Chapter 2. The remainder of our discussion will be devoted to nonparametric estimators based on the maximum likelihood criterion. A survey of nonparametric procedures may be found in Wegman [1972].

1.3 The Histogram

The histogram, the classical nonparametric probability density estimator, probably antedates any parametric estimator. Given a sample $(x_1,\ldots,x_N) \subset [a,b]^N$, we partition the interval $[a,b]$ by $a = t_1 < t_2 < \ldots < t_{m+1} = b$ and we consider all simple functions W defined on $[a,b]$ having the form

$$W(t) = y_i \quad \text{for} \quad t \in [t_i,t_{i+1}) \quad i = 1,m$$
$$W(b) = y_m \tag{1.3.1}$$

and zero elsewhere for some $(y_1,\ldots,y_m) \in R_+^m$. If we let q_i be the number of samples in the interval $[t_i,t_{i+1})$ for $i = 1,m$ and let q_m include the samples equal to b, then the histogram estimate is the simple function W defined by

$$y_i = \frac{q_i}{N(t_{i+1} - t_i)} \quad \text{for} \quad i = 1,m . \tag{1.3.2}$$

We first show that the histogram is the unique function of the form (1.3.1) which maximizes the likelihood function

$$L(W) = \prod_{i=1}^{N} W(x_i) \tag{1.3.3}$$

subject to the constraints

$$\int W(t)dt = 1 \quad \text{and} \quad W(t) \geq 0 \quad \forall\, t .$$

In intervals where $q_i = 0$, the optimal solution y_i^* must be zero, since any mass placed in the i^{th} interval decreases the likelihood.

Following de Montricher [1973], we prove a lemma and a proposition verifying that (1.3.2) is the unique solution to the constrained optimization problem (1.3.3). Some of these results may also be found in the paper by de Montricher, Tapia and Thompson [1975].

Lemma 1.3.1. Given positive integers $q_1,\ldots,q_m$ define $f:R^m \to R$ by

$$f(y) = \prod_{i=1}^{m} y_i^{q_i}$$

where $y = (y_1,\ldots,y_m)$. Also given $\alpha \in R^m$ such that $\alpha > 0$, define T by

$$T = \{y \in R^m: \langle \alpha, y \rangle = 1 \quad \text{and} \quad y \geq 0\}$$

where $\langle \cdot,\cdot \rangle$ denotes the usual inner product in R^m. Then f has a unique maximizer in T which is given by y^* where

$$y_i^* = \frac{q_i}{N\alpha_i} \quad \text{and} \quad \sum_{i=1}^{m} q_i = N .$$

Proof. Since T is compact and f is a continuous function of y, there exists a global maximizer which we denote by y^*. If y_i^* were zero, then $f(y^*) = 0$. But $y = \frac{1}{m}(\frac{1}{\alpha_1},\ldots,\frac{1}{\alpha_m}) \in T$ and $f(y) > 0$, which would be a contradiction. It follows that y^* must be an interior point of T. From the theory of Lagrange multipliers there exists $\lambda \in R$ such that

$$\nabla f(y^*) = \lambda\alpha . \tag{1.3.4}$$

Taking the gradient of f and using (1.3.4) leads to

$$q_i f(y^*) = \lambda\alpha_i y_i^* \qquad i = 1,m . \tag{1.3.5}$$

From (1.3.5) and the fact that $\langle \alpha, y^* \rangle = 1$ we have

$$\lambda = \lambda\langle \alpha, y^* \rangle = \sum_{i=1}^{m} f(y^*)q_i = Nf(y^*) .$$

Substituting this value for λ into (1.3.5) gives

$$q_i f(y^*) = Nf(y^*)\alpha_i y_i^*$$

establishing the lemma, since we have proved that (1.3.4) has a unique solution.

For the class of simple functions (1.3.1), the integral constraint is seen to be

$$\sum_{i=1}^{m} y_i(t_{i+1} - t_i) = 1$$

We may now prove the following:

Proposition 1.3.1. For a given partition, the histogram is the unique maximum likelihood estimator in the space of nonnegative simple functions of the form (1.3.1), given by

$$y_i = \frac{q_i}{N(t_{i+1} - t_i)} \qquad i = 1,m \tag{1.3.6}$$

where q_i and N are as before.

Proof. We have already noted that $y_i^* = 0$ in intervals where $q_i = 0$, and formula (1.3.6) is valid in this case. Let $I = \{i:q_i > 0\}$. Then applying lemma 1.3.1 over those indices in I with $\alpha_i = t_{i+1} - t_i$ completes the proof.

We next show that the histogram is a consistent estimator and we calculate the optimal rate of convergence for the histogram. The following proof is motivated by a proof for kernel estimators by Rosenblatt [1956], providing a link between histograms and kernel estimators.

Theorem 1.3.1. Suppose the sampling density $f(x)$ has continuous derivatives up to order three. Suppose we define a mesh on the real line, as described in (1.3.23), by the set $\{t_k\}$ for $-\infty < k < \infty$ where $t_{k+1} - t_k = h$

for all k and for a given mesh interval h. Then the histogram $W(t)$ defined as in (1.3.1) by

$$y_i = \frac{q_i}{Nh} \tag{1.3.7}$$

is consistent in mean square error in the sense that

$$E\left|W(x) - f(x)\right|^2 \to 0 \tag{1.3.8}$$

as $N \to \infty$, $h \to 0$, and $Nh \to \infty$. Furthermore, if $h = h(N)$ is chosen so that

$$h(N) = \left[\frac{2f(x)}{f'(x)^2}\right]^{1/3} N^{-1/3} \tag{1.3.9}$$

then the optimal rate of convergence of the mean square error is

$$E\left|W(x) - f(x)\right|^2 \sim 3\left[\frac{f(x)f'(x)}{\sqrt{2}}\right]^{2/3} N^{-2/3} + O(\tfrac{1}{N} + h^3) . \tag{1.3.10}$$

Proof. Let us consider the estimate at a fixed point x^*, where we assume that x^* is always in the k^{th} interval $[t_k, t_{k+1})$ as we change the mesh width h. Let us further suppose that the mesh is picked so that x^* is the midpoint of $[t_k, t_{k+1})$ even as we vary h. Then $t_{k+1} - x^* = \frac{1}{2}h$ and $x^* - t_k = \frac{1}{2}h$ for all h, where the dependence of the mesh nodes on h has been suppressed. Let

$$p_k = \int_{t_k}^{t_{k+1}} f(x)dx . \tag{1.3.11}$$

Taking a Taylor expansion of f about x^*, we get

$$f(x) = f(x^*) + f'(x^*)(x-x^*) + \tfrac{1}{2}f''(x^*)(x-x^*)^2 + O[(x-x^*)^3]. \tag{1.3.12}$$

Using (1.3.12) in (1.3.11) and noting the linear term drops out in the integral, we have

$$p_k = f(x^*)h + \frac{1}{24} f''(x^*)h^3 + O(h^4) \quad . \tag{1.3.13}$$

We separate the mean square error (1.3.8) into a variance and bias term by

$$E\left|W(x^*) - f(x^*)\right|^2 = \sigma^2(W(x^*)) + \left[E(W(x^*) - f(X^*))\right]^2 \quad . \tag{1.3.14}$$

Now for a given h, using (1.3.7), we have

$$E(W(x^*)) = \frac{1}{Nh} E(q_k) = \frac{p_k}{h} \tag{1.3.15}$$

and

$$\begin{aligned} \sigma^2(W(x^*)) &= E\left|\frac{q_k}{Nh} - \frac{p_k}{h}\right|^2 \\ &= \frac{1}{N^2h^2} E\left|q_k - Np_k\right|^2 \\ &= \frac{p_k(1 - p_k)}{Nh^2} \end{aligned} \tag{1.3.16}$$

where q_k is the number of samples in $[t_k, t_{k+1})$ and p_k is given by (1.3.11). Using (1.3.13) in (1.3.16) we get

$$\sigma^2(W(x^*)) = \frac{1}{Nh}\left[f(x^*) - f(x^*)^2h + \frac{1}{24} f''(x^*)h^2 + O(h^3)\right]. \tag{1.3.17}$$

Thus the variance of the histogram estimate will vanish as $N \to \infty$ and $h \to 0$ if we require that $Nh \to \infty$. Similarly we use (1.3.13), (1.3.14), and (1.3.15) to calculate the

$$(\text{bias})^2 = \left[\frac{p_k}{h} - f(x^*)\right]^2 = \frac{1}{24^2} f''(x^*)^2h^4 + O(h^5) \tag{1.3.18}$$

which vanishes as $h \to 0$. From (1.3.17) and (1.3.18) we have

$$\begin{aligned} E\left|W(x^*) - f(x^*)\right|^2 \sim \frac{f(x^*)}{Nh} + O\left(\frac{1}{N}\right) \\ + \frac{1}{24^2} f''(x^*)^2h^4 + O(h^5) \end{aligned} \tag{1.3.19}$$

where we have combined all the variance terms in (1.3.17) except the first under the term $O(1/N)$, since we will consider picking h of the form $N^{-\alpha}$ where $\alpha > 0$.

Let us consider the mean square estimate of the histogram at any other point y in the interval $[t_k, t_{k+1})$. Since $W(y) = W(x^*)$ and $(a+b)^2 \leq 2a^2 + 2b^2$ for any real numbers a and b, we have

$$E|W(y) - f(y)|^2 = E|W(x^*) - f(x^*) + f(x^*) - f(y)|^2$$
$$\leq 2E|W(x^*) - f(x^*)|^2 + 2E|f(x^*) - f(y)|^2 . \tag{1.3.20}$$

Now, using (1.3.12) with the worst value of y in $[t_k, t_{k+1})$, namely $y = t_k$, we have, since $x^* - t_k = h/2$, that

$$|f(x^*) - f(y)| \leq f'(x^*)\frac{h}{2} + o(h^2) . \tag{1.3.21}$$

Using (1.3.19) and (1.3.21) in (1.3.20) we obtain

$$E|W(y) - f(y)|^2 \leq \frac{2f(x^*)}{Nh} + \frac{f'(x^*)^2}{2} h^2 + o(\frac{1}{N} + h^3) . \tag{1.3.22}$$

The choice of $h(N)$ which minimizes the first two terms in (1.3.22) may be obtained directly or by using Lemma 4a in Parzen [1962, p. 1074] and is given by (1.3.9) with corresponding optimal mean square error (1.3.10).

Suppose we always halve h when we change h and that we define the new mesh by shifting any interval boundary in the old mesh by $h/4$. With this choice we always keep points that were midpoints of intervals in the previous mesh at midpoints of intervals in the new mesh. If we let $|$ denote an interval boundary and let $*$ denote the center of an interval, this algorithm looks like

```
  h : |       *       |       *       |
h/2 : *   |   *   |   *   |   *   |   *
h/4 : * | * | * | * | * | * | * | * | *
```
(1.3.23)

and so on. Clearly the points denoted by the asterisks become dense on the real line. This proves the theorem.

Corollary 1.3.1. Suppose zero is an interval midpoint in (1.3.23). Then under the conditions of Theorem 1.3.1 and the choice of

$$h(N) = \left[\frac{2^4 3^2 f(x^*)}{f''(x^*)^2}\right]^{1/5} N^{-1/5}$$

the optimal mean square error

$$E\left|W(x) - f(x)\right|^2 \sim 5\left[\frac{f''(x^*)^2 f(x^*)^4}{2^{14} 3^2}\right]^{1/5} N^{-4/5} + O\left(\frac{1}{N} + h^5\right)$$

is attained at any finite number of the points $x = kh$ for $k = 0, \pm 1, \pm 2, \ldots$ where x^* is chosen such that $f''(x)^2 f(x)^4$ is greatest. However, at other points in those intervals the mean square error may be as slowly decreasing $N^{-2/5}$.

Proof. Follows directly from (1.3.19), (1.3.23), and (1.3.22).

Corollary 1.3.2. Under the conditions of Theorem 1.3.1, the histogram is consistent in the integrated mean square error, that is,

$$E\int \left|W(x) - f(x)\right|^2 dx \to 0 \quad .$$

Furthermore, the choice

$$h(N) = \left[\frac{2}{\int f'(x)^2 dx}\right]^{1/3} N^{-1/3}$$

implies the optimal rate of convergence

$$E \int \left|W(x) - f(x)\right|^2 dx \sim 3\left[\tfrac{1}{2} \int f'(x)^2 dx\right]^{1/3} N^{-2/3} + O\left(\frac{1}{N} + h^3\right) .$$

Proof. Follows immediately after integrating (1.3.22).

The Histospline

A procedure based on the histogram is the histospline of Boneva, Kendall,

and Stefanov [1971]. Given an estimate of a histogram associated with a data set, the authors propose fitting a spline to the histogram in a manner which preserves areas in each mesh interval for the purpose of obtaining an estimate of the true density smoother than the histogram. It is shown that for an appropriate Hilbert space, there is a one-to-one correspondence between the histogram and the histospline. The authors argue that the presence of small negative values in the tails should not prove a serious problem. However, in practical classification schemes it is necessary to compare density values. No mention is made of how one might proceed if at least one of these values should prove negative. Another serious problem is the fact that the histospline introduces many local modes as the mesh interval width decreases. Thus the practitioner is forced to use wide intervals that may camouflage fine structure available in the data.

II. KERNEL ESTIMATORS

2.1 Description and Consistency of the Kernel Estimator

A fundamental theoretical advance in nonparametric density estimation beyond the counting estimates of frequency tables and histograms was made by Rosenblatt [1956]. He considered using a central difference of the sample cumulative distribution function as an estimate of the density, a form that Fix and Hodges [1951] had used in a nonparametric discrimination application. Rosenblatt proved that his estimate is asymptotically consistent in both mean square error and integrated mean square error. Rosenblatt's estimate is a member of the class of nonparametric density estimators that has come to be known as kernel estimators. However, it was Parzen [1962] who generalized and popularized the one-dimensional kernel estimator. His elegant treatment was generalized to multi-dimensional densities by Cacoullos [1966].

For p-dimensional data, a function $K(\cdot)$ is called a density kernel if the following conditions are satisfied:

$$K : R^p \to R^+ \tag{2.1.1}$$

$$K \in L^2(R^p)$$

$$\int_{R^p} K(y)dy = 1$$

$$\operatorname{ess\,sup}_{x \in R^p} K(x) < \infty$$

$$\lim_{\|x\| \to \infty} \|x\| K(x) = 0 \quad .$$

Let $K_h(y) = \frac{1}{h} K(\frac{y}{h})$. For a given random sample $x_1,\ldots,x_N$, the kernel estimator has the simple form

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^{N} K_h(y - x_i) \quad . \tag{2.1.2}$$

Clearly the condition that K be a density function insures that $\hat{f}$ will also be. We note that h is a scale parameter which reflects the spread or support of K_h. Furthermore, the estimate has equal weights of $1/N$ on each of the N kernels centered at the data points. That $h \to 0$ as $N \to \infty$ is an obvious requirement for consistency of the kernel estimator. From Bennett, de Figueiredo, and Thompson [1974] we have the following results concerning the consistency of the kernel estimator and the optimal rate of convergence.

Proposition 2.1.1. Suppose $x_1,\ldots,x_N$ is a random sample from $f(\cdot) \in L^2(R^p)$, $h = h(N)$ satisfies

$$\lim_{N\to\infty} h(N) = 0$$

$$\lim_{N\to\infty} Nh(N)^p = \infty \quad ,$$

and $K(\cdot)$ is a density kernel defined by conditions (2.1.1). Then the kernel estimate $\hat{f}(\cdot)$ defined by (2.1.2) is a consistent estimator of $f(\cdot)$ in the integrated mean square error; that is,

$$\lim_{N\to\infty} \|E(\hat{f}(x) - f(x))^2\|_1 = 0$$

where

$$\|\varphi\|_q = \left[\int_{R^p} |\varphi(x)|^q \, dx\right]^{1/q} \quad .$$

To obtain the optimal rate of convergence for a choice of K, we assume $f(\cdot)$ is three times Gateaux differentiable and that K is symmetric, i.e., $K(-x) = K(x)$. Then for the optimal choice

$$h(N) = \left[\frac{p\gamma_p}{\|I\|_2^2}\right]^{\frac{1}{p+4}} N^{-\frac{1}{p+4}} \tag{2.1.3}$$

where

$$\gamma_p = \|K\|_2^2$$

and

$$I = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \int_{R^p} y_i y_j K(y) dy$$

the optimal rate of convergence of the integrated mean square error is

$$(\frac{p}{4} + 1)^{\frac{-p}{p+4}} (\|I\|_2^2)^{\frac{p}{p+4}} \gamma_p^{\frac{4}{p+4}} N^{\frac{-4}{p+4}} + O(h^4) \quad . \tag{2.1.4}$$

For the one-dimensional case $p = 1$, we see that $h(N) \propto N^{-1/5}$ and that the error is of order $N^{-4/5}$. In order to use these expressions for $p = 1$, an estimate of the following is required;

$$\int_{-\infty}^{\infty} f''(x)^2 \, dx \quad . \tag{2.1.5}$$

In practical situations, however, prior knowledge of this quantity is rare. Fortunately, good estimates can be designed in an interactive mode, a procedure that is dealt with in some detail in Chapter 5 in connection with the penalized maximum likelihood algorithm. The idea is to pick h as small as possible without the variance of the resulting estimator becoming inconsistent with our prior feelings about the true density. This is an example of the bias-variance tradeoff that is well known in spectral analysis. For h too large, we have very smooth estimates, hence a small variance at the price of a large bias. For h too small, we may detect the fine structure observable from the data, but at the price of high variance. The point where the bias and variance of the estimate are both acceptable has largely been a subjective decision, best resolved in an interactive mode with the computer. In section 2.5 we consider a new, more objective method of choosing h.

The Fourier Kernel

If we relax any of the assumptions on the kernel in (2.1.1), we may hope to obtain significantly improved rates of convergence to zero of the mean square error for the kernel estimator. Davis [1975] considers the Fourier kernel

$$K(x) = \frac{\sin x}{\pi x} \tag{2.1.6}$$

which is neither nonnegative nor in $L^1(R)$, although it is in $L^2(R)$. Suppose the characteristic function $\varphi(\omega)$ of density $f(t)$ satisfies

$$|\varphi(\omega)| \leq Ae^{-\rho|\omega|^r} \tag{2.1.7}$$

for some constants $A > 0$, $\rho > 0$, and $0 < r \leq 2$, along with one other technical requirement. Then φ is said to decrease exponentially with degree r and coefficient ρ. The Normal and Cauchy densities are in this class with $\rho = \sigma^2/2$, $r = 2$ and $\rho = 1$, $r = 1$, respectively. When the kernel scaling factor $h(N)$ is chosen of the order $(\log \frac{n}{2\rho})^{-1/r}$ Davis shows that

$$\lim_{N\to\infty} \frac{\text{M.S.E. of the Fourier kernel}}{\text{M.S.E. of any (2.1.1) kernel}} = 0 \tag{2.1.8}$$

where M.S.E. denotes the mean square error of the estimate at some point.

If the characteristic function of $f(t)$ satisfies the weaker condition

$$\lim_{\omega\to\infty} |\omega|^q |\varphi(\omega)| > 0$$

then the characteristic function is said to have algebraic decrease of order q. This class includes the chi-squared and exponential densities with $q = \frac{1}{2}$(degrees of freedom) and $q = 1$, respectively. The optimal choice for $h(N)$ is of order $N^{-q/2}$, and Davis shows (2.1.8) holds for $q > 5/2$.

In practical terms, the Fourier kernel introduces negative estimates

as does the histospline discussed in section 1.3, resulting in the same ambiguities. To use the optimal results, we need to have strong prior knowledge about the characteristic function of the unknown density. This requirement is much more stringent than, say, prior knowledge of (2.1.5). The small sample properties of the Fourier kernel are not evident in the above discussion. In chapter 5, we demonstrate the undesirable small sample properties of this estimator.

2.2 An Optimal Kernel

Whittle [1958] attacked the problem of finding an optimal kernel for estimating the density at a point, based on prior information about the density, without knowledge of Rosenblatt's work. In section 2.3 we show that in a sense Parzen's kernel estimator is a special case of Whittle's estimator when there is no prior information available.

Epanechnikov [1969] observed that the expression for the optimal rate of convergence of the integrated mean square error (2.1.4) had two factors involving the kernel:

$$\int K^2(x)dx$$

and (2.2.1)

$$\int x^2 K(x)dx$$

where we consider the one-dimensional case $p = 1$. Following Rosenblatt [1971], this leads one to consider the optimization problem

$$\begin{aligned} &\text{minimize} && \int K^2(x)dx && (2.2.2)\\ &\text{subject to} && \int K(x)dx = 1 \\ & && K(-x) = K(x) \geq 0 \\ & && \int x^2 K(x)dx = 1 \,. \end{aligned}$$

In a short variational argument the optimal kernel is calculated to be

$$K(x) = \begin{cases} 3/4\ 5^{-\frac{1}{2}}(1 - x^2/5) & \text{if } |x| \le \sqrt{5} \\ 0 & \text{otherwise} \end{cases} . \qquad (2.2.3)$$

This is a nonnegative function with finite support. Bennett, de Figueiredo, and Thompson [1974] chose a B-spline for the kernel function partly because of this property. Philosophically, kernels with finite support seem attractive on the grounds that the resulting density has zero mass in the tails. Only when theoretical considerations have lead to a specific parametric density should we feel confident about estimates in the tails outside the range of the data.

A criticism of Epanechnikov's kernel is that it attempts to minimize the bound on the error, which may not be a sharp bound. We see that this kernel is minimax in flavor, trying to minimize the worst that might happen. A generalization of Epanechnikov's work may be found in Kazakos [1975].

We consider the problem complementary to (2.2.2) where we reverse the roles of the two factors (2.2.1) involving the kernel in the optimal error expression (2.1.4):

$$\begin{aligned} \text{minimize} \quad & \int x^2 K(x)dx \\ \text{subject to} \quad & \int K(x)dx = 1 \\ & K(-x) = K(x) \ge 0 \\ & \int K^2(x)dx = 1 \quad . \end{aligned}$$

It is a straightforward exercise to verify that the kernel solving this problem is

$$K(x) = \frac{5}{4}\left(1 - \frac{25}{9}x^2\right) \quad \text{for} \quad |x| \le \frac{3}{5} \qquad (2.2.4)$$

and zero elsewhere. If we scale x by a factor $5\sqrt{5}/3$, (2.2.4) is identical to the kernel (2.2.3) obtained in the original problem (2.2.2). It should be mentioned that using kernels with finite support has definite computational advantages.

2.3 An Optimally Smooth Kernel Estimate

Whittle [1958] considered finding estimates of a density function at a point x in an optimal fashion using a kernel estimator. His kernel denoted by $\omega_x(\cdot)$ depends on x. His estimate takes the form

$$\hat{f}(x) = \frac{1}{N} \sum_{j=1}^{N} \omega_x(x_j) \ . \tag{2.3.1}$$

Whittle assumes that the number of observations N is a Poisson variable with mean M. The kernel is chosen to minimize the expected mean square error

$$\Delta^2 = E_p E_S[\hat{f}(x) - f(x)]^2 \tag{2.3.2}$$

where E_S denotes expectation with respect to the random sampling, and E_p denotes expectation with respect to the prior distribution of the ordinates of the unknown density function. In particular, functions $\mu(x)$ and $\mu(x,y)$ are assumed known *a priori* such that

$$\begin{aligned} E_p[Mf(x)] &= \mu(x) \\ E_p[Mf(x)Mf(y)] &= \mu(x,y) \end{aligned} \tag{2.3.3}$$

where M is the Poisson mean described above. Then the optimal kernel $\omega_x(y)$ minimizing (2.3.2) solves the integral equation

$$\mu(y)\omega_x(y) + \int \mu(y,z)\omega_x(z)dz = \mu(y,x) \ . \tag{2.3.4}$$

Whittle notes that $\omega_x(y) \sim \delta(y-x)$ for large expected sample sizes and that $\omega_x(y)$ is invariant to scalings of the density function. In this general case for a given sample he can demonstrate neither that his estimate is nonnegative nor that his estimate applied everywhere integrates to one.

For convenience Whittle defines the normalized kernel $\xi_x(y)$ and normalized covariance function $\gamma(x,y)$ by

$$\xi_x(y) = \omega_x(y)\sqrt{\frac{\mu(y)}{\mu(x)}}$$

$$\gamma(x,y) = \frac{\mu(x,y)}{\sqrt{\mu(x)\mu(y)}} \qquad (2.3.5)$$

so that equation (2.3.4) becomes

$$\xi_x(y) + \int\gamma(y,z)\xi_x(z)dz = \gamma(y,x) \ . \qquad (2.3.6)$$

As a special case he considers a normalized covariance function that is second-order stationary, in keeping with the time series flavor of his approach. Whittle notes this assumption is plausible if the prior $\mu(\cdot)$ is a diffuse uniform density. Replacing $\gamma(x,y)$ with $K(y-x)$ according to the second-order assumption, the optimal kernel $\xi_x(y)$ must satisfy the integral equation

$$\xi_x(y) + \int_a^b K(y-z)\xi_x(z)dz = K(y-x) \ . \qquad (2.3.7)$$

We have the following proposition:

Proposition 2.3.1. Suppose $K(\cdot)$ satisfies

$$\int_a^b\int_a^b |K(y-x)|^2 dxdy < \infty \qquad (2.3.8)$$

where the interval (a,b) may be infinite. Then equation (2.3.7) has a unique solution $\xi_x(y)$ in $L^2(-\infty,\infty)$.

Proof. If (2.3.8) is satisfied, then the operator $T(\cdot)$ defined by $T(\eta) = \int_a^b K(y-z)\eta(z)dz$ is a Hilbert-Schmidt compact operator. Thus equation (2.3.7) may be written in the form $(I+T)\xi = K$, an equation for which solutions exist and are unique.

Assuming $(a,b) = (-\infty,\infty)$, Whittle takes the Fourier transform twice in equation (2.3.7) to solve for the optimal normalized kernel as

$$\xi_x(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{i\omega(y-x)}k(\omega)}{1 + k(\omega)}\, d\omega \qquad (2.3.9)$$

where $k(\omega)$ is the Fourier transform of $K(x)$. Whittle has made use of the convolution theorem to solve for (2.3.9). However, it is customary (see Stein [1971], p. 3) to assume that both $K(\cdot)$ and $\xi_x(\cdot)$ are in $L^1(-\infty,\infty)$ for the Fourier transform of the convolution to exist. As a particular choice of $K(\cdot)$, Whittle considers

$$K_\alpha(x) = \nu(\alpha + \beta e^{-\gamma|x|}) \tag{2.3.10}$$

where ν is the average density of observations and α, β, and γ are non-negative constants. Clearly, for $\alpha \neq 0$, $K(\cdot)$ is not an $L^1(-\infty,\infty)$ function. However, using (2.3.9) Whittle "solves" for the optimal normalized kernel with the result

$$\xi_x(y) = \frac{\nu\beta\gamma}{\theta} e^{-\theta|y-x|} \tag{2.3.11}$$

where

$$\theta = (2\nu\beta\gamma + \gamma^2)^{\frac{1}{2}} . \tag{2.3.12}$$

We note (as Whittle does) that $\xi_x(y)$ does not depend on the choice of α. Therefore, solving the problem for α and α' should lead to the same solution $\xi_x(y)$. Substituting $\xi_x(y)$ into equation (2.3.7) for the values α and α' and subtracting implies

$$\int [K_\alpha(y-z) - K_{\alpha'}(y-z)]\xi_x(z)dz = K_\alpha(y-x) - K_{\alpha'}(y-x)$$

or

$$\int (\nu\alpha - \nu\alpha')\xi_x(z)dz = \nu\alpha - \nu\alpha' . \tag{2.3.13}$$

We have the following:

Theorem 2.3.1. A necessary condition for $\xi_x(\cdot)$ to be the solution of the integral equation (2.3.7) with covariance kernel (2.3.10) for arbitrary α is that

$$\int \xi_x(z)dz = 1 . \tag{2.3.14}$$

Proof. The proof follows immediately from (2.3.13).

We may take (2.3.14) as a constraint that must be satisfied by the solution to (2.3.7). If a solution happened to satisfy the constraint for a particular choice of α, β, and γ, then clearly it would not for a slightly perturbed value of γ or β. For the particular solution (2.3.11) to problem (2.3.7) with covariance kernel (2.3.10) this integral condition is easily seen to be

$$\frac{2\nu\beta\gamma}{\theta^2} = 1 .$$

Using the definition of θ (2.3.12), we have immediately that $\gamma^2 = 0$ which in turn implies $\theta = 0$. Thus we have shown

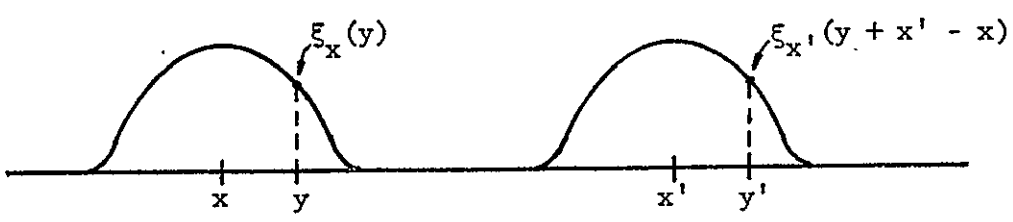
Theorem 2.3.2. Equation (2.3.11) is never a solution to problem (2.3.7) with covariance kernel (2.3.10) for arbitrary α.

Proof. The solution (2.3.11) is undefined for $\gamma = 0$.

If we consider $\alpha = 0$, then $K_0(\cdot)$ is an $L^1(-\infty,\infty)$ function and the solution given is correct, as may be verified by direct substitution.

Let us see how the second-order stationarity assumption allows us to view Whittle's estimator as a Parzen estimator, that is, to estimate the entire density with one kernel.

We claim that $\xi_x(y) = \xi_0(y-x)$, that is, the functions $\xi_x(\cdot)$ and $\xi_{x'}(\cdot)$ solving equation (2.3.7) for any x and x' are identical in form.



This is equivalent to saying:

Theorem 2.3.3. The optimal normalized kernels solving (2.3.7) under assumption (2.3.8) on the real line at any two points x and x' satisfy

$$\xi_x(y) = \xi_{x'}(y + x' - x) . \tag{2.3.15}$$

Proof. Recall that $\xi_x(y)$ uniquely satisfies by Proposition 2.3.1

$$\xi_x(y) + \int_{-\infty}^{\infty} K(y-z)\xi_x(z)dz = K(y-x) \quad \forall\ y .$$

Making the change of variable $y \to \omega + x - x'$ which has Jacobian of unity, we obtain

$$\xi_x(\omega+x-x') + \int_{-\infty}^{\infty} K(\omega+x-x'-z)\xi_x(z)dz = K(\omega-x') .$$

Transforming again $z \to v + x - x'$, we obtain

$$\xi_x(\omega+x-x') + \int_{-\infty}^{\infty} K(\omega-v)\xi_x(v+x-x')dv = K(\omega-x') .$$

Replacing $\omega \to y$, $v \to z$ and exchanging x and x', we finally have

$$\xi_{x'}(y+x'-x) + \int_{-\infty}^{\infty} K(y-z)\xi_{x'}(z+x'-x)dz = K(y-x) .$$

As a function of y $\xi_{x'}(y+x'-x)$ solves equation (2.3.7). Since $\xi_x(y)$ was the unique solution of (2.3.7), we must have (2.3.15), verifying our claim that $\xi_x(y) = \xi_0(y-x)$, choosing $x' = 0$.

Now let us suppose that $\mu(x)$ is rectangular on a very large interval as Whittle describes as sufficient for the second-order stationarity assumption. Approximately, we suppose $\frac{\mu(x)}{\mu(y)} = 1$ for all x,y. Then by (2.3.5) the unnormalized kernel $\omega_x(y)$ is identical with the normalized kernel $\xi_x(y)$ and Whittle's estimate becomes

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \xi_0(x-x_i)$$

which is the form of a Parzen estimator. For his particular choice of $K_\alpha(\cdot)$ with $\alpha = 0$, the estimate is

$$\hat{f}(x) = \frac{\nu\beta\gamma}{N\theta} \sum_{i=1}^{N} e^{-\theta|x-x_j|}$$

which can be normalized for a particular sample so that $\int_{-\infty}^{\infty} \hat{f}(x) = 1$; the nonnegativity of $\hat{f}(x)$ is obvious.

In summary, we see that in the case of second-order stationarity of the prior covariance function with constant (i.e., "informationless") prior ordinate values, the optimal Whittle estimator takes the form of the Parzen estimator. The generalized Picard kernel (see Davis [1975], p. 1026) results from the particular choice of $K(x) = \nu\beta e^{-\gamma|x|}$ for a fixed sample.

2.4 Unequal Weights for the Kernel Estimator

A natural question to be answered concerning the Parzen estimator is whether it can be improved. We consider allowing the estimator to place unequal weights on the kernels. We choose these weights to maximize the likelihood function (1.3.3). Specifically, for a choice of kernel $K_h(\cdot)$ we find the weights $\alpha_1,\ldots,\alpha_N$ that solve the following constrained optimization problem:

$$\text{maximize} \quad \prod_{i=1}^{N} \hat{f}(x_i) \tag{2.4.1}$$

$$\text{where} \quad \hat{f}(y) = \sum_{i=1}^{N} \alpha_i K_h(y-x_i) \tag{2.4.2}$$

$$\text{subject to} \quad \alpha_i \geq 0 \quad i = 1,N$$

$$\sum_{i=1}^{N} \alpha_i = 1 \quad . \tag{2.4.3}$$

The first constraint guarantees that $\hat{f}(y) \geq 0$, while the second constraint assures us that $\hat{f}(\cdot)$ integrates to one.

Following de Montricher [1973], we first establish the existence of the estimator (2.4.2) solving (2.4.1).

Proposition 2.4.1. If φ_i $i = 1,N$ is a set of linearly independent probability densities on $(-\infty,\infty)$, then there exists a function W of the form

$$W(t) = \sum_{i=1}^{N} \alpha_i \varphi_i(t)$$

which maximizes the likelihood $\prod_{i=1}^{N} W(x_i)$ among all functions of this form subject to the constraints $\sum_{i=1}^{N} \alpha_i = 1$ and $\alpha_i \geq 0$ for $i = 1,N$.

Proof. Let β_j denote the value of W at x_j and let $\beta = (\beta_1,\ldots,\beta_N)^t$. Therefore, by definition,

$$\beta_j = \sum_{i=1}^{N} \alpha_i \varphi_i(x_j) \quad . \tag{2.4.4}$$

If we define a square matrix A by its components $A_{ji} = \varphi_i(x_j)$, $i,j = 1,N$, then equation (2.4.4) becomes $\beta = A\alpha$, where $\alpha = (\alpha_1,\ldots,\alpha_N)^t$. This shows that β depends continuously on α; hence, the likelihood is a continuous function of α. Clearly the constraint set for α is a compact set in R^N. This proves the proposition.

Proposition 2.4.2. If K is a nonzero function in $L^2(-\infty,\infty)$ and $\{x_i\}$ $i = 1,N$ is a set of N distinct points on the real line, then the functions $\{K(t-x_i)\}$ $i = 1,N$ are linearly independent.

In view of the kernels advocated by Epanechnikov [1969] and Bennett [1974], it is of interest to consider the case where K has finite support and is a probability density. Although kernels are generally assumed to be in $L^2(-\infty,\infty)$, for kernels of finite support in $L^p(a,b)$, we have the following result.

Proposition 2.4.3. If K is a probability density function with finite support on an interval (a,b) and $\{x_i\}$ $i = 1,N$ is a set of N distinct points on the real line, then the functions $\{K(t-x_i)\}$ $i = 1,N$ are linearly independent.

Proof. We assume without loss of generality that $x_1 < x_2 < \ldots < x_N$ and that (a,b) is symmetric about the origin, that is, $a = -b$. Suppose $\psi(t) = \sum_{i=1}^{N} \alpha_i K(t-x_i) = 0$ in $L^p(-\infty,\infty)$. On the interval (x_1-b, x_2-b) we see that $\psi(t) = \alpha_1 K(t-x_1)$, since this is the only term where the kernel is nonzero. This implies that $\alpha_1 = 0$ in order that $\psi(t) = 0$ in the L^p sense on the interval (x_1-b, x_2-b), which has positive Lebesgue measure by assumption. Continuing this reasoning inductively, we conclude that $\alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$, proving the proposition.

Remark. For a continuous density function, the assumption that the random samples $x_1,\ldots,x_N$ are distinct holds with probability one.

To study the uniqueness of the weights in (2.4.2), we begin by looking at the convexity properties of the likelihood $\prod_{i=1}^{N} W(x_i)$.

Proposition 2.4.4. The functional $\varphi: R_+^N \to R$ defined by $\varphi(\beta) = \prod_{i=1}^{N} \beta_i$, where $\beta = (\beta_1,\ldots,\beta_N)^t$ and $\beta_i > 0$ for $i = 1,N$, has at most one maximizer over any convex subset of R_+^N.

Proof. Suppose $\varphi(\beta) = \varphi(\bar{\beta}) = C$. Then

$$\sum_{i=1}^{N} \log \beta_i = \sum_{i=1}^{N} \log \bar{\beta}_i = \log C .$$

Let $E(\theta) = \varphi(\theta\varphi + (1-\theta)\bar{\beta})$ for $\theta \in (0,1)$. Then using the strict concavity of the log, we have

$$\log E(\theta) > \theta \sum_{i=1}^{N} \log \beta_i + (1-\theta) \sum_{i=1}^{N} \log \bar{\beta}_i = \log C .$$

Since the log is strictly increasing, $E(\theta) > C$ for $\theta \in (0,1)$. This proves the proposition since two distinct maximizers would lead to a contradiction.

Proposition 2.4.5. Any two solutions to the maximum likelihood problem stated in Proposition 2.4.1 coincide at the sample points $\{x_i\}$ $i = 1,N$.

Proof. Utilizing the notation introduced in previous propositions, the likelihood can be written as

$$J(\alpha) = \varphi(A(\alpha)) .$$

Let T denote the constraint set for α as in proposition 2.4.1. Maximizing J subject to the constraint $\alpha \in T$ is equivalent to maximizing φ over $A(T)$. It follows that $A(T)$ is convex and compact. The likelihood φ is continuous; hence it has a maximizer say $\bar{\beta}$ over $A(T)$. Moreover, this solution is unique by Proposition 2.4.4. It follows that the set of all solutions to the original problem (2.4.1) is $\{\alpha \in T | A\alpha = \bar{\beta}\}$. This proves the proposition.

De Montricher [1973] demonstrates a kernel where (2.4.1) does not have a unique solution. Thus the matrix A is not invertible in general. We answer de Montricher's question of uniqueness by making the following stochastic statement which gives sufficient conditions for A to be invertible for certain kernels:

Theorem 2.4.1. Suppose that the kernel K is nonnegative, symmetric, and strictly positive at the origin, that is, $K(x) = K(|x|)$ and $K(0) > 0$. Suppose that the sampling density is absolutely continuous. Furthermore, assume that for all x such that $K(x) \neq 0$, $\mu\{y:K(y) = K(x)\} = 0$, where μ denotes Lebesgue measure. Then with probability one, A is invertible and problem (2.4.1) has a unique solution for a fixed sample size N.

Proof. (By induction). Let $k_o = K(0)$. For $N = 1$ using the notation of the previous propositions, the matrix $A_1 = [k_o]$, where the subscript on A specifically denotes N . Thus A_1 is nonsingular since $K(0) \neq 0$ by hypothesis.

For $N = 2$

$$A_2 = \begin{bmatrix} k_o & K(\Delta_{12}) \\ K(\Delta_{12}) & k_o \end{bmatrix}$$

where $\Delta_{ij} = |x_i - x_j|$. A_2 is nonsingular if $K(\Delta_{12}) \neq k_o$ which occurs with probability one. Now suppose after N samples A_N is nonsingular. We take another sample x_{N+1} . We inquire about the nonsingularity of

$$A_{N+1} = \left[\begin{array}{c|c} A_N & b \\ \hline b^t & c \end{array}\right]$$

where $A_{ij} = A_{ji} = K(\Delta_{ij})$, $b_i = K(\Delta_{i,N+1})$, and $c = K(\Delta_{N+1,N+1}) = k_o$, for $i,j = 1,N$. Does there exist an $(N+1) \times 1$ vector $v = (\gamma^t | \alpha)^t \neq 0$, where γ is an $N \times 1$ vector and α is a constant such that $A_{N+1}v = 0$? Suppose so; now $A_{N+1}V = 0$ is equivalent to the pair of equations

$$A_N\gamma + b\alpha = 0 \qquad (2.4.5)$$

$$b^t\gamma + c\alpha = 0 \quad . \qquad (2.4.6)$$

First suppose $\alpha = 0$. Then $A_N\gamma = 0$ which in turn implies $\gamma = 0$, since A_N is nonsingular by the induction hypothesis. But $\alpha = 0$ and $\gamma = 0$ gives $v = 0$, contradicting our assumption that A_{N+1} is singular. We note that if $x_i = x_j$, then two rows of A_{N+1} would be identical and A_{N+1} singular.

Now for $\alpha \neq 0$ we have from equation (2.4.5) that

$$\gamma = -A_N^{-1}b\alpha \quad .$$

Substituting this value for γ into equation (2.4.6), cancelling an α factor and noting that $c = k_o$, we obtain an equation for b

$$b^t A_N^{-1} b = k_o \neq 0 \quad . \tag{2.4.7}$$

Solutions of equation (2.4.7) determine the vector v and lie on a surface in R^N known as a "central quadric" (see Noble [1969], p. 391) since A_N is symmetric. Clearly $b = 0$ is not a solution of equation (2.4.7). Since the sampling density is absolutely continuous by hypothesis and K does not assume the same nonzero value on a set of positive measure, the probability that x_{N+1} results in a vector b satisfying (2.4.7) is zero. Therefore, with probability one A_{N+1} is nonsingular, proving the proposition.

Preliminary Numerical Results

We may now consider the effect of optimizing the likelihood of the Parzen estimator with unequal weights. No theoretical conclusions have as yet been made, so several computer simulations motivate the following. Using the quartic kernel K_3 given in Table 2.5.1 on several random samples of size 25 from the standard Gaussian distribution, we immediately see that most of the weights are set equal to zero. In fact, no more than four weights were nonzero in our few trials. The kernel scaling parameter h was chosen as the best value for the usual Parzen estimator. The finite support characteristic of the kernel seemed to play a dominant role in the resulting kernal weight estimates. Sample points outside the range of a single large kernel weight near the mean required a small weight to get a positive likelihood. The resulting estimate retains the character of the particular kernel, but involves few kernel evaluations. Some fruitful research should be possible in this area. An algorithm similar to the one

presented in Chapter 5 was used to calculate the optimal weights.

2.5 A Data-Oriented Procedure for Picking the Best h(N) Value for a Sample from an Unknown Density

For a univariate density we know from (2.1.3) that the asymptotically optimal choice of $h(N)$ for a given kernel $K(y)$ in the Parzen estimator of $f_o(x)$ is of the form

$$h(N) = \alpha(K)\beta(f_o)N^{-1/5} \tag{2.5.1}$$

where

$$\alpha(K) = \left\{\frac{\int K(x)^2 dx}{\left[\int x^2 K(x)dx\right]^2}\right\}^{1/5} \tag{2.5.2}$$

and

$$\beta(f_o) = \left[\int f_o''(x)^2 dx\right]^{-1/5} . \tag{2.5.3}$$

In this section we consider a procedure based only on the samples $x_1,\ldots,x_N$ for choosing a value of $h(N)$ without any prior knowledge of $\beta(f_o)$.

We first remark that equation (2.5.1) gives the asymptotic optimal choice for $h(N)$ that minimizes the integrated mean square error. For small sample sizes, (2.5.1) gives the choice of $h(N)$ that is optimal only on the average over all possible random samples of size N . Suppose we know the true density $f(x)$. Then for a given sample of size N and kernel $K(y)$, there is a value $h^*(N)$ that actually minimizes the integrated mean square error

$$\int_{-\infty}^{\infty} \left[\frac{1}{N}\sum_{i=1}^{N} K_h(y-x_i) - f_o(y)\right]^2 f_o(y)dy . \tag{2.5.4}$$

This value $h^*(N)$ which we call the best choice will be close to the optimal value given by (2.5.1).

Suppose for a given sample we have a good estimate of $h(N)$ with

the kernel K_1, call it h_1. We propose using (2.5.1) to get a good estimate of $h(N)$ for any other kernel K_2, call it h_2, by scaling h_1 as follows:

$$h_2 = \frac{\alpha(K_2)}{\alpha(K_1)} h_1 \quad . \tag{2.5.5}$$

We shall present a procedure for finding a good estimate for the Normal kernel and then argue that an application of formula (2.5.5) will give a good estimate for any other kernel satisfying (2.1.1). In order to present empirical evidence of (2.5.5), we introduce four kernels. The first three have their support on the interval $[-1,1]$.

TABLE 2.5.1

$K(\cdot)$		$\alpha(K)$
$K_1(y) = \frac{1}{2}$	$\lvert y\rvert \le 1$	1.3510
$K_2(y) = 1 - \lvert y\rvert$	$\lvert y\rvert \le 1$	1.8882
$K_3(y) = \frac{15}{16}(y^4+1) - \frac{15}{8} y^2$	$\lvert y\rvert \le 1$	2.0362
$K_4(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$	$\lvert y\rvert < \infty$	0.7764

The first two kernels are the box and triangle, respectively. The third kernel is a quartic polynomial with coefficients chosen so that $K(y)$ and $K'(y)$ vanish at $y = \pm 1$ with the usual integral constraint. In practice, K_3 results in estimates virtually indistinguishable from the Gaussian kernel K_4. Tremendous computational advantages result from using the finite support of K_3 and avoiding exponentials.

Three data sets were generated. For several of the kernels given in Table 2.5.1, the best choice of $h(N)$ was computed; a line search was used

to find that value h^* minimizing a Simpson's rule approximation to the integrated mean square error (2.5.4). The first data set was a sample of size 10 from the standard normal density. The following table summarizes the values h_i^* obtained by minimizing (2.5.4) with kernel K_i. The extrapolated estimates h_j were obtained using (2.5.5):

TABLE 2.5.2. Sample of size 10 from N(0,1)

Best h_i value	Good h_i value extrapolated by (2.5.5) from		
	h_1^*	h_2^*	h_3^*
$h_1^* = 1.1561$	--	1.21	1.20
$h_2^* = 1.6933$	1.62	--	1.68
$h_3^* = 1.8117$	1.74	1.83	--

A second data set was a sample of size 100, also from the N(0,1).

TABLE 2.5.3. Sample of size 100 from N(0,1)

Best h_i value	Good h_i value extrapolated by (2.5.5) from		
	h_1^*	h_2^*	h_3^*
$h_1^* = 0.6890$	--	0.75	0.74
$h_2^* = 1.0478$	0.96	--	1.04
$h_3^* = 1.1189$	1.04	1.13	--

The third data set of size 100 was from the bimodal mixture density $\frac{1}{2}\,N(-1.5,1) + \frac{1}{2}\,N(1.5,1)$.

TABLE 2.5.4. Sample of size 100 from mixture density

Best h_i value	Good h_i value extrapolated from	
	h_3^*	h_4^*
$h_3^* = 1.0500$	--	1.09
$h_4^* = 0.4148$	0.40	--

Thus application of formula (2.5.5) is seen to give estimates of $h^*(N)$ for the other kernels with a relative error of less than 9%.

It may be shown that for the Parzen estimator $\hat{f}$ based on the Normal kernal K_4, given a sample $x_1,\ldots,x_N$ and any $h > 0$

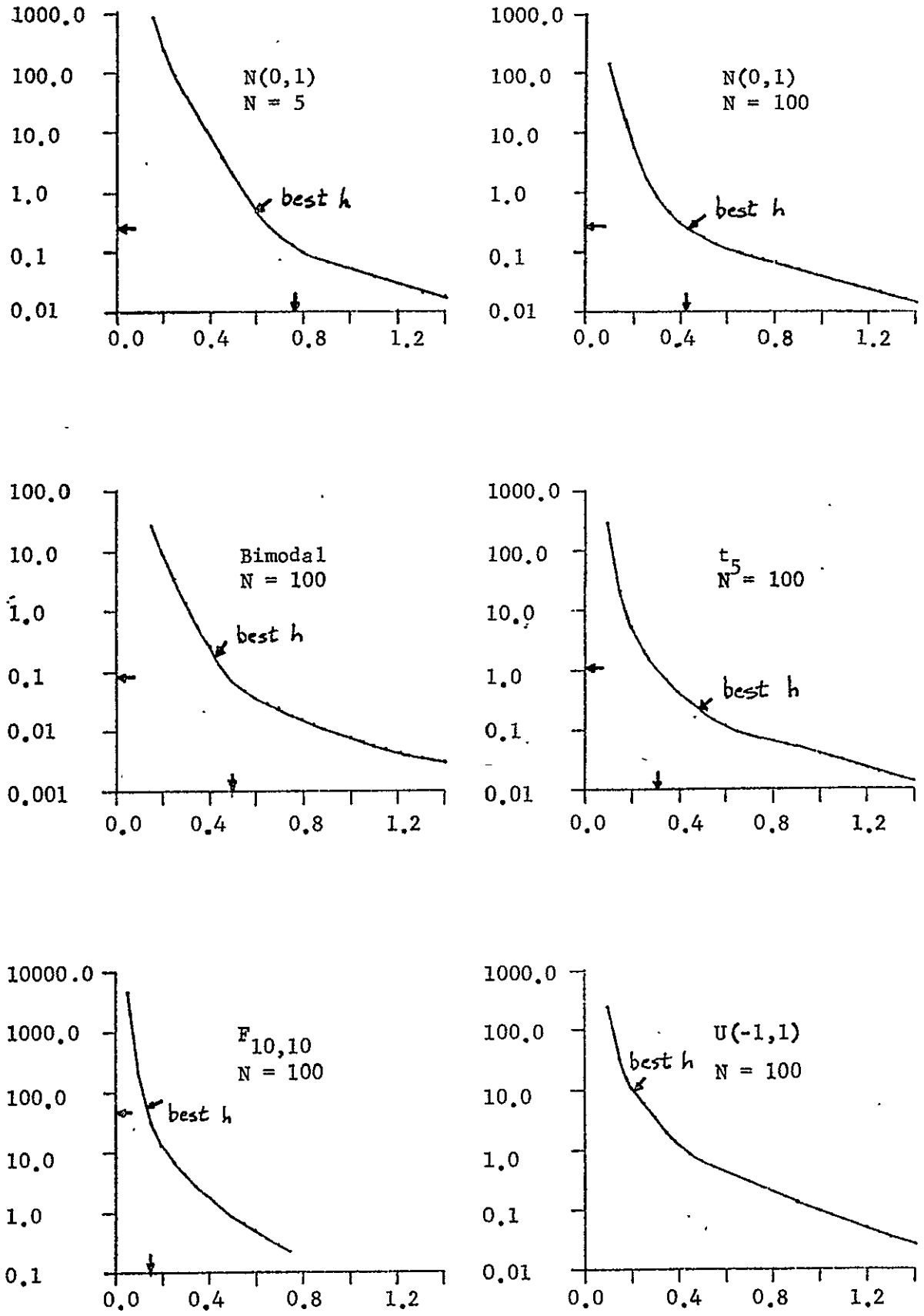
$$\int_{-\infty}^{\infty} \hat{f}''(y)^2 dy = \frac{3}{8\sqrt{\pi}\, N^2 h^9} \sum_{j=1}^{N} \sum_{k=1}^{N} \left[h^4 - (x_j - x_k)^2 + \frac{1}{12}(x_j - x_k)^4\right] e^{-(x_j-x_k)^2/4h^2} . \qquad (2.5.6)$$

The proposed procedure is to plot (2.5.6) for a range of values of $h > 0$. Empirical evidence suggests that (2.5.6) is nearly zero for large values of h. On the other hand, as h approaches zero, (2.5.6) blows up at least as fast as an exponential. Intuitively, as h decreases from a large value, the increase in (2.5.6) is due to an improvement in the approximation of the Parzen estimator to the true density. As h approaches zero, the rapid increase in (2.5.6) results as the Parzen estimator begins to resemble a linear combination of equally weighted Dirac spikes. Between these two extremes the best h for the data lies. Finally, an application of (2.5.5) will give a good estimate for any other kernel.

Remark. For the best h, (2.5.6) will generally be greater than the similar quantity associated with the true density found in (2.5.3). This difference is due to the small variations in the best Parzen estimate not found in the true density.

As empirical evidence we consider six samples from a total of five densities. In Figure 2.5.1 the computed best h from the line search algorithm is marked on the curve in each graph. The theoretical values of $h(N)$ and $\int f_o''(x)^2 dx$ are marked on the axes where applicable. We plot

FIGURE 2.5.1. Graphs of Equation (2.5.6) vs. h



the results on semi-log paper to accommodate the wide variation in the computed values of (2.5.6).

In section 2.1 we discussed the interactive mode for choosing h in light of the bias-variance tradeoff. Using graphs such as those in Figure 2.5.1, the process of decreasing h until the variance of the estimate is unacceptable may be performed. On the semi-log scale we consider the curve in three portions. In the first portion, the quantity (2.5.6) for the estimate increases in an approximately linear fashion as h decreases. In the next portion of the curve, which looks like a heel, the quantity (2.5.6) increases more rapidly. For h's in the heel, the fine structure of the true density becomes apparent in the corresponding estimate. Finally, above the heel as an h approaches zero, the quantity (2.5.6) becomes infinite with the Dirac spike estimate. Choosing h at the beginning of exponential part of the heel is recommended in view of the remark preceding the last paragraph.

2.6 A Quasi-Optimal Procedure

We propose an algorithm based on functional iteration to calculate automatically the best h for a random sample from an unknown density. Let $h^{(0)}$ be an initial guess. Let $\beta(h)$ denote the quantity (2.5.3) corresponding to (2.5.6) for the Parzen estimate with that choice of h :

$$\beta(h) = \left[\int_{-\infty}^{\infty} \hat{f}''(y)^2 dy \right]^{-1/5} \qquad (2.6.1)$$

Using the Gaussian kernel and (2.5.1), consider

$$h^{(1)} = (2\sqrt{\pi}\, N)^{-1/5}\, \beta(h^{(0)})$$

or in general

$$h^{(i+1)} = q(h^{(i)}) \qquad (2.6.2)$$

letting

$$q(h^{(i)}) = (2\sqrt{\pi}\,N)^{-1/5}\,\beta(h^{(i)}) \tag{2.6.3}$$

where the superscript on h denotes the current iteration number. If $h^{(i+1)} = h^{(i)}$ in (2.6.2), we define this value of h to be a solution. Thus we are looking for nonnegative values of h where the two functions

$$\begin{aligned} \varphi_1(h) &= h \\ \varphi_2(h) &= q(h) \end{aligned} \tag{2.6.4}$$

agree. For a given sample $x_1,\ldots,x_N$ we may graph (2.6.4) for all h using (2.6.3) and (2.5.6) and observe where the functions intersect. To examine the behavior of (2.6.4), we examine (2.5.6). For large values of h, (2.5.6) is approximately

$$\frac{3}{8\sqrt{\pi}\,N^2h^9}\sum_{j=1}^{N}\sum_{k=1}^{N}h^4 = \frac{3}{8\sqrt{\pi}\,h^5}$$

or

$$\beta(h) \approx \left(\frac{8\sqrt{\pi}}{3}\right)^{1/5} \cdot h \quad \text{as } h \to \infty \ . \tag{2.6.5}$$

As $h \to 0$, (2.5.6) $\to \infty$; hence,

$$\beta(h=0) = 0 \ . \tag{2.6.6}$$

Therefore, using (2.6.1), (2.6.3), (2.6.5) and (2.6.6), we have

$$q(h=0) = 0 \tag{2.6.7}$$

and

$$q(h) \approx \left(\frac{4}{3N}\right)^{1/5} \cdot h \quad \text{as} \quad h \to \infty \ . \tag{2.6.8}$$

At $h = 0$ the functions (2.6.4) agree. We call $h = 0$ the <u>degenerate solution</u>. From (2.6.8) for $N > 1$ we see that $q(h)$ is approximately linear with slope less than one. Thus the functions (2.6.4) can agree only for small values of h. We call the <u>largest</u> value of h where the functions (2.6.4) agree the <u>quasi-optimal</u> value of h. Clearly this value

FIGURE 2.6.1. Plots of Functions (2.6.4) vs. h
For Several Samples

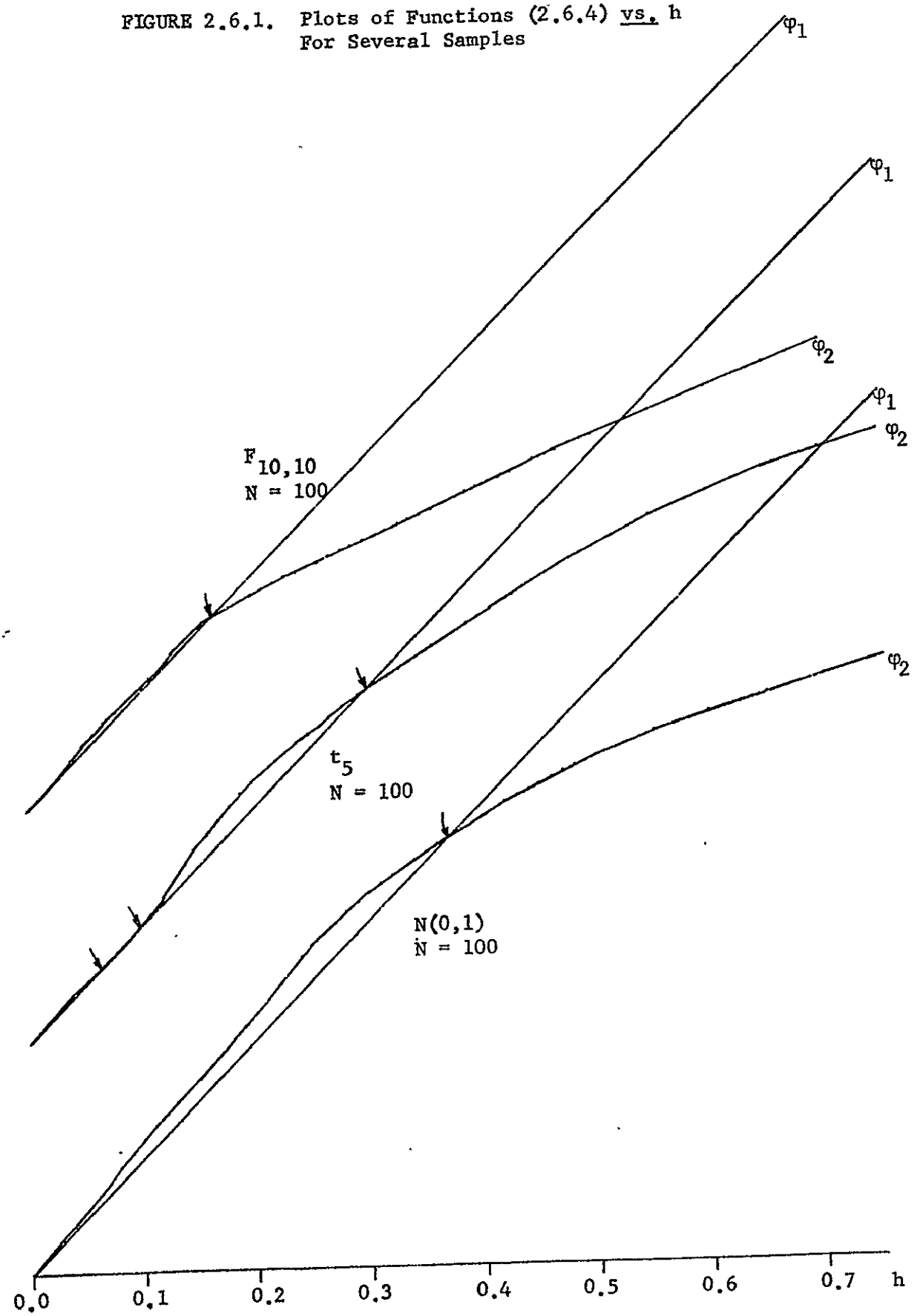
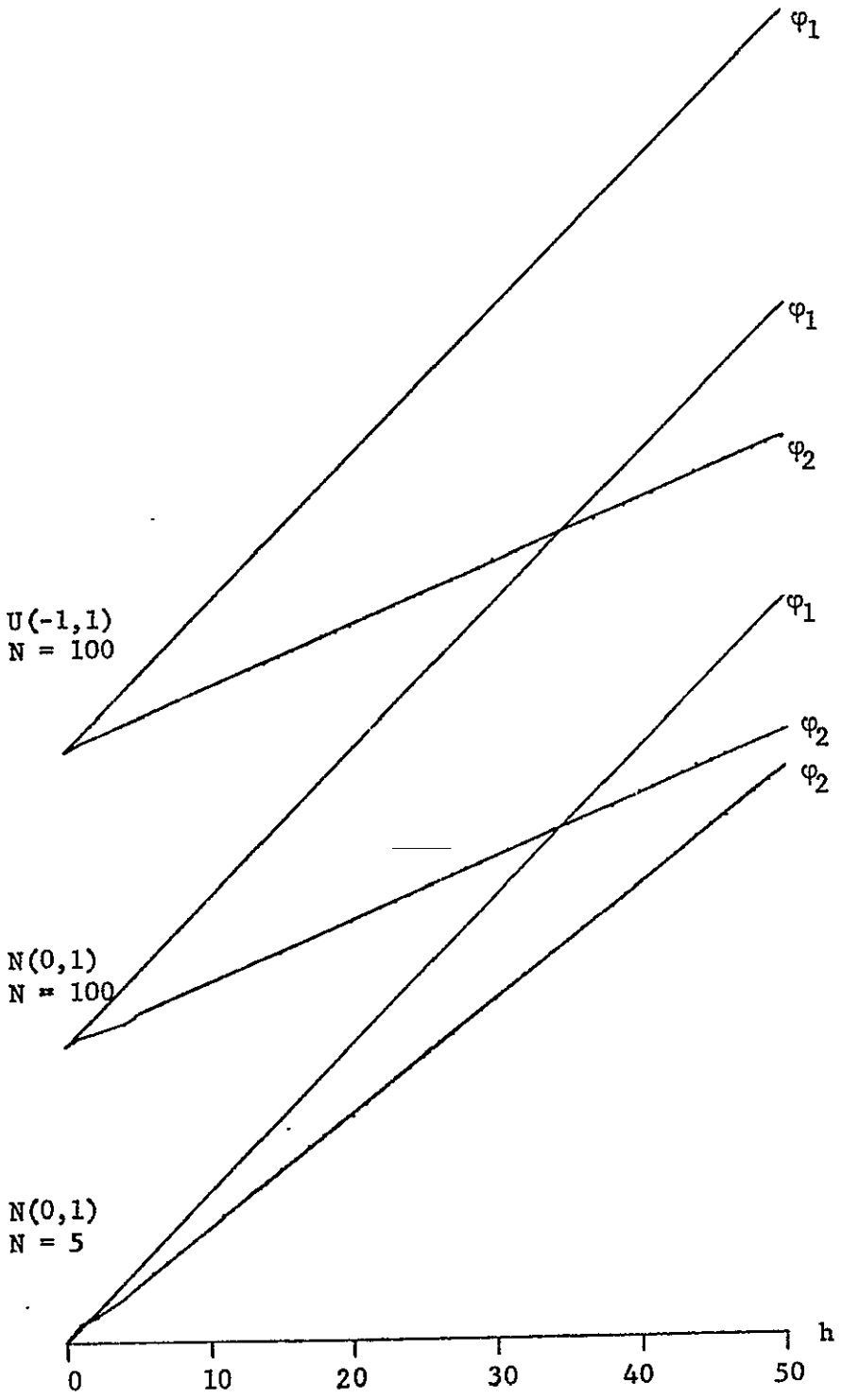


FIGURE 2.6.2. Plots of Functions (2.6.4) vs. h For Several Samples



exists and is unique, although it may be degenerate.

In Figures 2.6.1 and 2.6.2 we graph the functions (2.6.4) for several random samples. In Figure 2.6.1 we see that several solutions may exist but the quasi-optimal solution is unique. The nondegenerate solutions are marked on these graphs. In Figure 2.6.2 the asymptotic behavior predicted by (2.6.8) is clearly evident.

The functional iteration algorithm (2.6.2) is ideal for finding the quasi-optimal solution without the possibility of converging to another solution. We simply pick $h^{(0)}$ large enough so that we are on the linear portion of $q(h)$. The iterates look like:

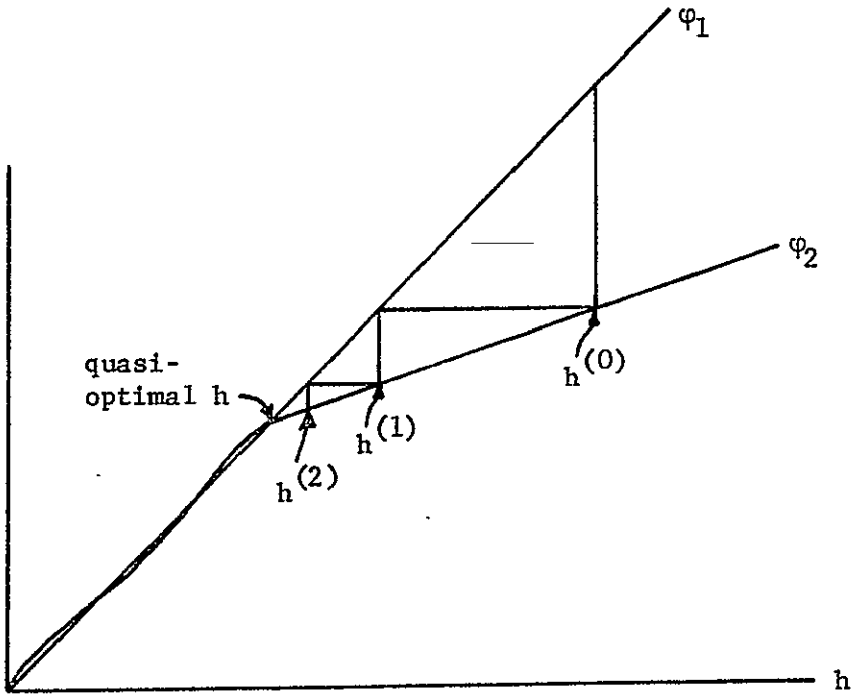


FIGURE 2.6.3

Convergence is fast away from the solution. A Newton's method step may be inserted alternately to speed up convergence near the solution, rejecting the Newton step if it is "too big." A necessary condition for h^* to be quasi-optimal is $|q'(h^*)| < 1$; that is, the functional iteration scheme will converge to h^* only if this condition is met.

Several Monte Carlo studies were performed using the quasi-optimal algorithm. Six densities were tried (the Bimodal form is given before Table 2.5.4 and the Cauchy density has scale parameter one). Twenty-five random samples of size 25 and 100 were generated for each of the six densities. In all cases, $h^{(0)}$ was taken to be one. Of the 400 samples generated, 370 converged using Newton's method alone, usually in four iterations. The necessary condition $|q'(h)| < 1$ was verified. Of the remaining 30 samples, functional iteration indicated 12 were degenerate. A solution was accepted as quasi-optimal if $|h-q(h)| < 10^{-5}$. In Table 2.6.1 we summarize the results of the Monte Carlo study. The mean, standard deviation, and range of the calculated (nondegenerate) quasi-optimal solutions are given along with the theoretically optimal value given by (2.1.3).

For the quasi-optimal and the theoretically optimal choices of h , the integrated mean square error was estimated for the samples generated in Table 2.6.1. The calculation was performed using Simpson's rule on the interval (-5,5) with mesh spacing of a tenth. We remark that for the $F_{10,10}$ density, the interval was (-2,8) . In Table 2.6.2 we summarize the results of this exercise. The mean and standard deviation of the estimated integrated mean square error for the quasi-optimal and theoretically optimal choices of h are given. The efficiency is simply the ratio of the two estimated means. In Chapter 5, with these same samples, a sensitivity study is summarized. The integrated mean square error is calculated for

those h's differing from the optimal h by a factor of two. The efficiency for these choices of h is generally worse than the efficiency of the quasi-optimal choice.

TABLE 2.6.1. Monte Carlo Results for Finding the Quasi-Optimal h (Each row represents 25 samples.)

Density	Sample Size	Degenerate	Mean	Std. Dev.	Range	Theoretical
N(0,1)	25	1	.54	.17	.20 - .80	.56
Bimodal	25	2	.77	.41	.09 - 1.35	.66
Cauchy	25	1	.65	.25	.26 - 1.09	.54
U(-1,1)	25	1	.21	.12	.02 - .39	--
t_5	25	0	.59	.19	.25 - .96	.41
$F_{10,10}$	25	2	.25	.09	.02 - .42	.20
N(0,1)	100	0	.35	.10	.09 - .51	.42
Bimodal	100	0	.43	.17	.12 - .76	.50
Cauchy	100	0	.37	.12	.16 - .62	.41
U(-1,1)	100	4	.14	.04	.07 - .23	--
t_5	100	0	.37	.09	.13 - .54	.31
$F_{10,10}$	100	1	.15	.04	.05 - .20	.15

TABLE 2.6.2. Integrated Mean Square Error Using the Quasi-Optimal h *vs.* the Theoretically Optimal h

Density	Number of Samples	Sample Size	Quasi-Optimal h		Theoretical h		Efficiency
			Mean	Std. Dev.	Mean	Std. Dev.	
N(0,1)	24	25	.0066	.0057	.0043	.0032	65%
Bimodal	22*	25	.0037	.0053	.0014	.0011	38%
t_5	25	25	.0056	.0031	.0048	.0023	86%
$F_{10,10}$	22*	25	.0172	.0120	.0157	.0106	91%
N(0,1)	24*	100	.0019	.0013	.0013	.0008	68%
Bimodal	23*	100	.0008	.0004	.0005	.0003	63%
t_5	25	100	.0021	.0028	.0016	.0010	76%
$F_{10,10}$	24	100	.0091	.0112	.0067	.0052	74%

* The largest one or two I.M.S.E. values were omitted because the corresponding quasi-optimal h was nearly zero. The values omitted were Bimodal 25 (.0248), $F_{10,10}$ 25 (.8049), N(0,1) 100 (.0178), and Bimodal 100 (.0055, .0056).

III. NONPARAMETRIC MAXIMUM LIKELIHOOD DENSITY ESTIMATORS

3.1 Introduction

Following the success of the maximum likelihood philosophy in the parametric density estimation case it was only natural that attempts would be made to employ the principle of maximum likelihood in the nonparametric case. Given a random sample $x_1,\ldots,x_N$ from a density function defined on the set $\Omega = (a,b)$, we define the likelihood that a function $f \in L^1(\Omega)$ gave rise to the random sample as

$$L(f) = \prod_{i=1}^{N} f(x_i) \quad . \tag{3.1.1}$$

If we pick a manifold $H(\Omega) \subset L^1(\Omega)$, we may consider the following constrained optimization problem:

$$\begin{aligned} &\text{maximize} && L(f) \\ &\text{subject to} && f \in H(\Omega) \ , \ \int f(t)dt = 1 \ , \\ &\text{and} && f(t) \geq 0 \ \ \forall \ t \in \Omega \ . \end{aligned} \tag{3.1.2}$$

The integration is with respect to Lebesgue measure. Any solution to problem (3.1.2) is defined to be a maximum likelihood estimate based on the sample $x_1,\ldots,x_N$. As discussed in Chapter 1, unless a specific functional form for the density is assumed in $H(\Omega)$, we shall refer to all solutions of problem (3.1.2) as nonparametric. Perhaps a further distinction is required based on the dimensionality of the manifold $H(\Omega)$. We shall mainly be considering infinite dimensional manifolds and approximations of such manifolds.

The difficulty with problem (3.1.2) as stated is that a linear combination of Dirac delta functions at the sample points satisfies the constraints and results in a value of $+\infty$ for the objective likelihood functional. Clearly, while this estimate is representative of our sample, it

does not reflect the true population density. Unfortunately, most infinite dimensional manifolds can approximate delta functions. De Montricher, Tapia, and Thompson [1975] note that continuous functions, differentiable functions, infinitely differentiable functions, and polynomials enjoy this approximation property. Thus for these choices of $H(\Omega)$, the maximum likelihood estimate does not exist. For finite dimensional manifolds $H(\Omega)$ we may observe poor robustness, i.e., we may be unable to approximate a wide range of potential "true" densities.

One solution to our dilemma is to pick a finite dimensional manifold in a very judicious manner. We have seen in Section 1.3 that the histogram is the maximum likelihood estimate for $H(\Omega)$ given by (1.3.1). Furthermore, the histogram enjoys the property of consistency. In section 2.4 we considered another example of a maximum likelihood estimate based on unequal weights in the kernel estimator.

Wegman, in a series of papers [1969, 1970, 1976] considered a maximum likelihood estimate similar to the histogram. His class of admissible estimators $H(\Omega)$ is the simple functions. The unusual feature of his $H(\Omega)$ is that the mesh is determined by the samples themselves. In the earlier works, the estimate was taken to be upper semi-continuous with mesh nodes at each sample point. This estimate proved to peak dramatically at the mode. Consequently, he was forced to assume prior knowledge of the location of the mode, introducing a modification that avoided the problem. In the most recent paper he observes that for a fixed number of nodes the optimal placement of the nodes with respect to the maximum likelihood criterion is at the sample points. Clearly the estimate is zero outside the sample range $[x_1,x_N]$. Finally, demanding that for m intervals there be at least k points in the closure of each interval, he proves density consistency

as $N \to \infty$ for appropriate rates of increase for m and k as functions of N. In this manner Wegman avoids the problem of peaking at the mode.

A second solution to the problem of guaranteeing existence for problem (3.1.2) was introduced in 1971 by Good and Gaskins [1972]. The authors, in fact, did not prove existence of solutions. In 1973, de Montricher, et al. [1975] proved both existence and uniqueness of solutions to the problem posed by Good and Gaskins. In the following sections we discuss the problem and extend the results to cases of practical interest.

3.2 The Maximum Penalized Likelihood Estimate

To avoid the difficulty of delta function candidates in problem (3.1.2), Good and Gaskins [1972] suggested formulating a penalty functional $\Phi: H(\Omega) \to R_+$ which would evaluate the smoothness of a particular density estimate on an interval scale. Here by the notion of smoothness we do not mean f has many continuous derivatives. Rather, we wish to avoid rapid oscillatory behavior in the estimate. To satisfy the nonnegativity constraint the authors chose the sometimes dangerous trick of working with $\sqrt{f}$. De Montricher, Tapia, and Thompson [1975], whose results we shall present below, prove that working with $\sqrt{f}$ is not always equivalent to working with f in the presence of a nonnegativity constraint.

We now define the Φ-penalized likelihood of $f \in H(\Omega)$ by

$$\hat{L}(f) = \prod_{i=1}^{N} f(x_i) \exp(-\Phi(f)) \tag{3.2.1}$$

for a given sample $x_1, \ldots, x_N$. Consider the constrained optimization problem

$$\begin{aligned} &\text{maximize} \quad \hat{L}(f) \\ &\text{subject to} \quad f \in H(\Omega) \ , \ \int_{\Omega} f(t)dt = 1 \ , \end{aligned} \tag{3.2.2}$$

and $f(t) \geq 0 \quad \forall\, t \in \Omega$.

Any solution of (3.2.2) is called a maximum penalized likelihood estimate (M.P.L.E.).

The structure available with a Hilbert Space makes it a natural choice for $H(\Omega)$. In particular we have the notion of orthogonality available with the inner product. The norm induced by the inner product is given by $\|f\|^2 = \langle f,f \rangle$ for $f \in H(\Omega)$. If $f_n \to f$ in $H(\Omega)$ implies $f_n(x) \to f(x) \quad \forall\, x \in \Omega$, then point evaluation is a continuous operation. In this case we say that $H(\Omega)$ is a reproducing kernel Hilbert space (R.K.H.S.).

We shall require that $H(\Omega)$ contain at least one feasible solution f for any $x_1,\ldots,x_N$ where $x_i \in \Omega$. Then there exists $f \in H(\Omega)$ such that

$$\int_\Omega f(t)dt = 1 \ , \ f(t) \geq 0 \quad \forall\, t \in \Omega \ ,$$

and (3.2.3)

$$f(x_i) > 0 \quad \text{for} \quad i = 1,N \ .$$

The following theorem from de Montricher [1975] gives sufficient conditions so that solutions to problem (3.2.2) exist and are unique.

Theorem 3.2.1. Suppose that $H(\Omega)$ is a R.K.H.S., integration over Ω is a continuous functional and there exists at least one $f \in H(\Omega)$ satisfying (3.2.3). Then the maximum penalized likelihood estimate exists and is unique.

Proof. The constraints in (3.2.2) form a closed convex subset of $\{f \in H(\Omega): f(x_i) > 0 \ , \ i = 1,N\}$. It may be shown that the penalized likelihood function is bounded from above. Combined with the weak compactness of the unit ball in $H(\Omega)$ the above results lead to the existence of a maximizer. The second Frechet derivative of the log $\hat{L}(f)$ is negative

definite. Hence, $\hat{L}(f)$ is strictly concave and has at most one maximizer on a convex set.

In section 3.3 we will motivate penalty functionals Φ involving integrals of various derivatives of f squared. Thus a natural choice for $H(\Omega)$ is a Sobolev space of order s denoted by $H^s(a,b)$. If s is an integer, then $f \in H^s(a,b)$ if and only if $f, f^{(1)},\ldots,f^{(s)} \in L^2(a,b)$ and the norm is given by

$$\|f\|^2_{H^s(a,b)} = \sum_{i=0}^{s} \alpha_i \|f^{(i)}\|^2_{L^2(a,b)} \tag{3.2.4}$$

where $\alpha_i \geq 0$ and $\alpha_0,\alpha_s > 0$. The interval (a,b) may be infinite in the above. If the interval is finite, we define f to be zero outside (a,b). To apply Theorem 3.2.1, we need the following lemma:

Lemma 3.2.1. The Sobolev space $H^s(a,b)$ is a R.K.H.S. if and only if $s > \frac{1}{2}$. In such a case, integration on (a,b) is a continuous linear functional if and only if (a,b) is a finite interval.

Remark. Theorem 3.2.1 and Lemma 3.2.1 imply that the M.P.L.E. corresponding to $H(\Omega) = H^s(a,b)$ where $s > \frac{1}{2}$ and (a,b) is a finite interval is well defined. If we define for integer values of s and finite interval (a,b)

$$H_0^s(a,b) = \{f \in H^s(a,b): f^{(i)}(a) = f^{(i)}(b) = 0,\ i = 0,s-1\} \tag{3.2.5}$$

we may show the M.P.L.E. is well defined for

$$H_0^s(a,b) \quad \text{where} \quad \|f\|^2 = \int_a^b f^{(s)}(t)^2 dt \quad . \tag{3.2.6}$$

We note that the norm in (3.2.6) is a natural choice for the penalty functional $\Phi(f)$ in (3.2.2). For an infinite interval, we have to consider norms such as

$$\|f\|^2 = \sum_{i=0}^{s} \int_{-\infty}^{\infty} \mu(t) f^{(i)}(t)^2 dt \qquad (3.2.7)$$

in order to satisfy the conditions of Theorem 3.2.1. Here $\mu(t) \geq c_1 + c_2 t^2$ for $-\infty < t < \infty$ with $c_1, c_2 > 0$. Thus if we wish to consider penalty functionals of the form (3.2.6) on the entire real line we are apparently forced into (3.2.7). In section 3.4 we show that this is not the case.

Remark. Consistency of the M.P.L.E. on the infinite interval has not yet been established. A straightforward argument by Good and Gaskins [1975] shows that for any $f \neq f_o$, the true density, $E[\hat{L}(f_o)] > E[\hat{L}(f)]$ holds for a sample size N large enough. Unfortunately, to establish consistency one must prove that a sequence of solutions to (3.2.2) for increasing sample size N converges.

We conclude this section with a theorem [de Montricher, et al., 1975] that characterizes the M.P.L.E. on a finite interval with (3.2.6) as the penalty functional Φ.

Theorem 3.2.2. The maximum penalized likelihood estimate corresponding to the Hilbert space $H_0(a,b)$ exists, is unique and is a polynomial spline (monospline) of degree 2s. Moreover, if the estimate is positive in the interior of an interval, then in this interval it is a polynomial spline of degree $2s$ and continuity class $2s-2$ with knots exactly at the sample points.

3.3 An Estimator of Good and Gaskins

Good and Gaskins [1972] considered the penalty functional

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)}\, dt \qquad (\alpha > 0) \,. \qquad (3.3.1)$$

This form arises since they chose to work with $\sqrt{f}$ in place of f to avoid the nonnegativity constraint. The penalty functional (3.3.1) is seen to be equivalent to

$$\Phi(f) = 4\alpha \int_{-\infty}^{\infty} \left[\frac{d\sqrt{f(t)}}{dt}\right]^2 dt \quad . \tag{3.3.2}$$

Clearly $\sqrt{f} \in H^1(-\infty,\infty)$ is the correct choice for $H(\Omega)$. De Montricher, Tapia, and Thompson [1975] proved that solutions corresponding to (3.3.1) are well defined. After some analysis they also demonstrated the unique solution to (3.2.2) with penalty function (3.3.1) by

$$\begin{aligned} f(t) &= g_\lambda(t)^2 \\ g_\lambda(t) &= \tfrac{1}{2} \sum_{i=1}^{N} \frac{v(t-x_i)}{g_\lambda(x_i)} \end{aligned} \tag{3.3.3}$$

where

$$v(t) = \frac{1}{2\sqrt{2\alpha\lambda}}\, e^{-\sqrt{\lambda/2\alpha}\,|t|} \qquad -\infty < t < \infty$$

and $\lambda > 0$ is a Lagrange multiplier.

We may calculate the exact solution (3.3.3) to Good's problem by simply calculating the N values $g_\lambda(x_i)$, $i = 1,N$. We have done this by picking the values x_k for t in equation (3.3.3) and arriving at the N nonlinear equations

$$g_\lambda(x_k) = \tfrac{1}{2} \sum_{i=1}^{N} \frac{v(x_k-x_i)}{g_\lambda(x_i)} \qquad k = 1,N \quad . \tag{3.3.4}$$

The Lagrange multiplier λ is picked so that

$$\int_{-\infty}^{\infty} f(t)dt = \int_{-\infty}^{\infty} g_\lambda(t)^2 dt = 1$$

This choice of λ is unique. Without loss of generality, we may assume that $x_1 \le x_2 \le \ldots \le x_N$. We may calculate

$$16\lambda\sqrt{2\alpha\lambda}\int_{-\infty}^{\infty} g_\lambda(t)^2 dt = \sum_{i=1}^{N} \frac{1}{g_\lambda(x_i)^2}$$

$$+ \sum_{i<j}\sum \frac{1}{g_\lambda(x_i)g_\lambda(x_j)} \left\{ \exp\left(-\sqrt{\frac{\lambda}{2\alpha}}\,(x_i+x_j)\right) \right.$$

$$\exp\left(\sqrt{\frac{2\lambda}{\alpha}}\,x_i\right) + \sqrt{\frac{2\lambda}{\alpha}}(x_j-x_i)\,\exp\left(-\sqrt{\frac{\lambda}{2\alpha}}\,(x_j-x_i)\right)$$

$$\left. + \exp\left(\sqrt{\frac{\lambda}{2\alpha}}\,(x_i+x_j)\right)\,\exp\left(-\sqrt{\frac{2\lambda}{\alpha}}\,x_j\right)\right\} \quad . \qquad (3.3.5)$$

For a given value of λ, Newton's method was used to solve the equations (3.3.4) for $g_\lambda(x_k)$. Since the integral (3.3.5) is infinite for $\lambda = 0$, zero for $\lambda = \infty$, and monotone decreasing inbetween, a simple line search was required to find λ^* such that the integral constraint was satisfied. Two examples are given in Figures 3.3.1 and 3.3.2.

We may observe the effects of the finite dimensional approximation to problem (3.3.2) employed by Good and Gaskins. They considered a Hermite function expansion for the solution retaining no more than the first fifty terms. The exponential nature of the curve is smoothed at the sample points. With a penalty functional (3.3.2) involving the second derivative squared, solutions have a greater fullness, a property that the Hermite functions seem to display. However, no exact solution is known for the latter penalty functional.

Remark. We see that working with $\sqrt{f}$ introduces $f(x)$ in the denominator of the penalty functional (3.3.1). Thus where $f(x)$ is large, the weight on the penalty function is reduced. This explains why Good and Gaskins' estimate tends to peak at the sample points. If we work with f itself and consider penalty functionals like (3.2.6), we may avoid this undesirable feature.

FIGURE 3.3.1. N = 3 N(0,1) Exponential Spline Solution

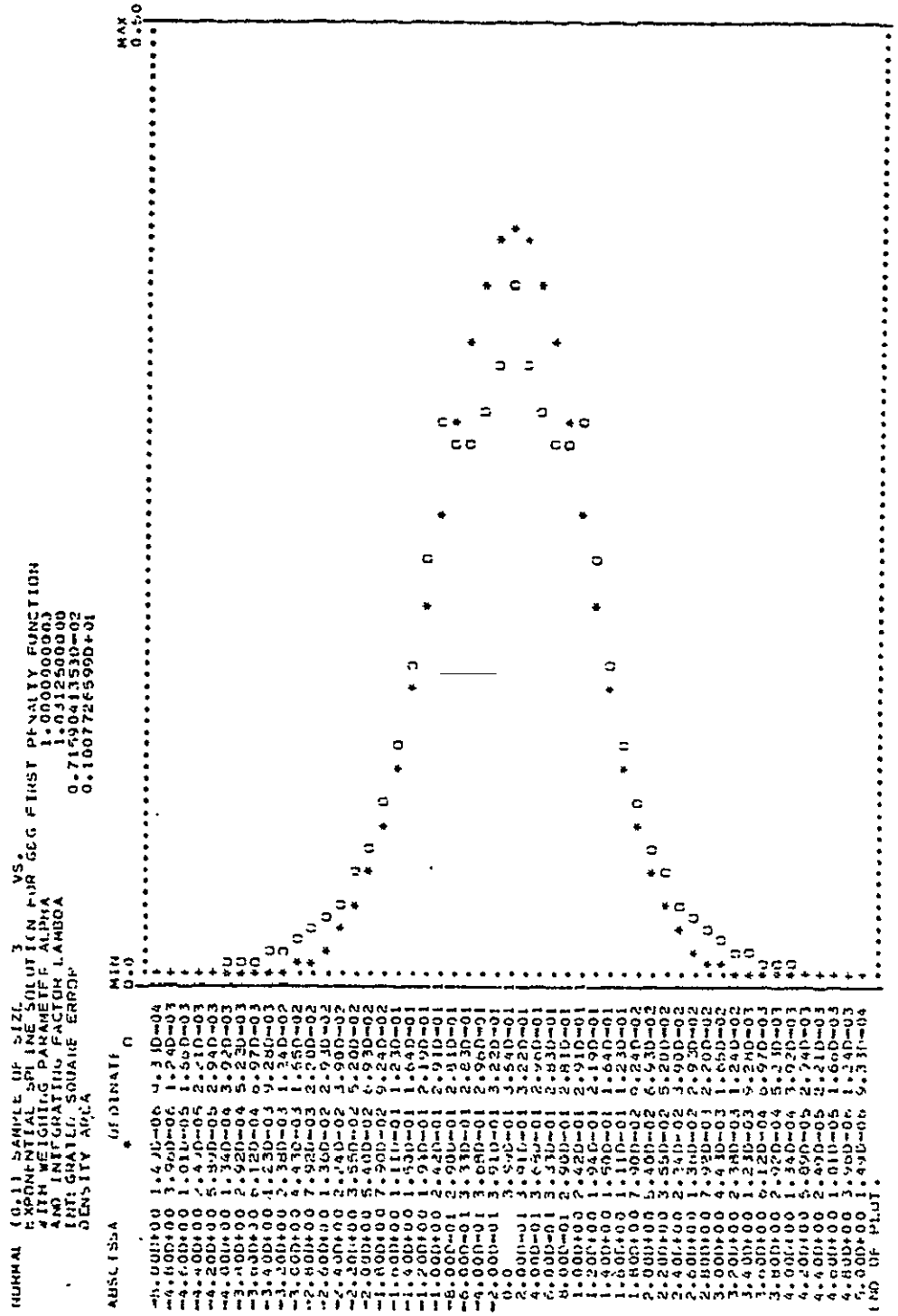
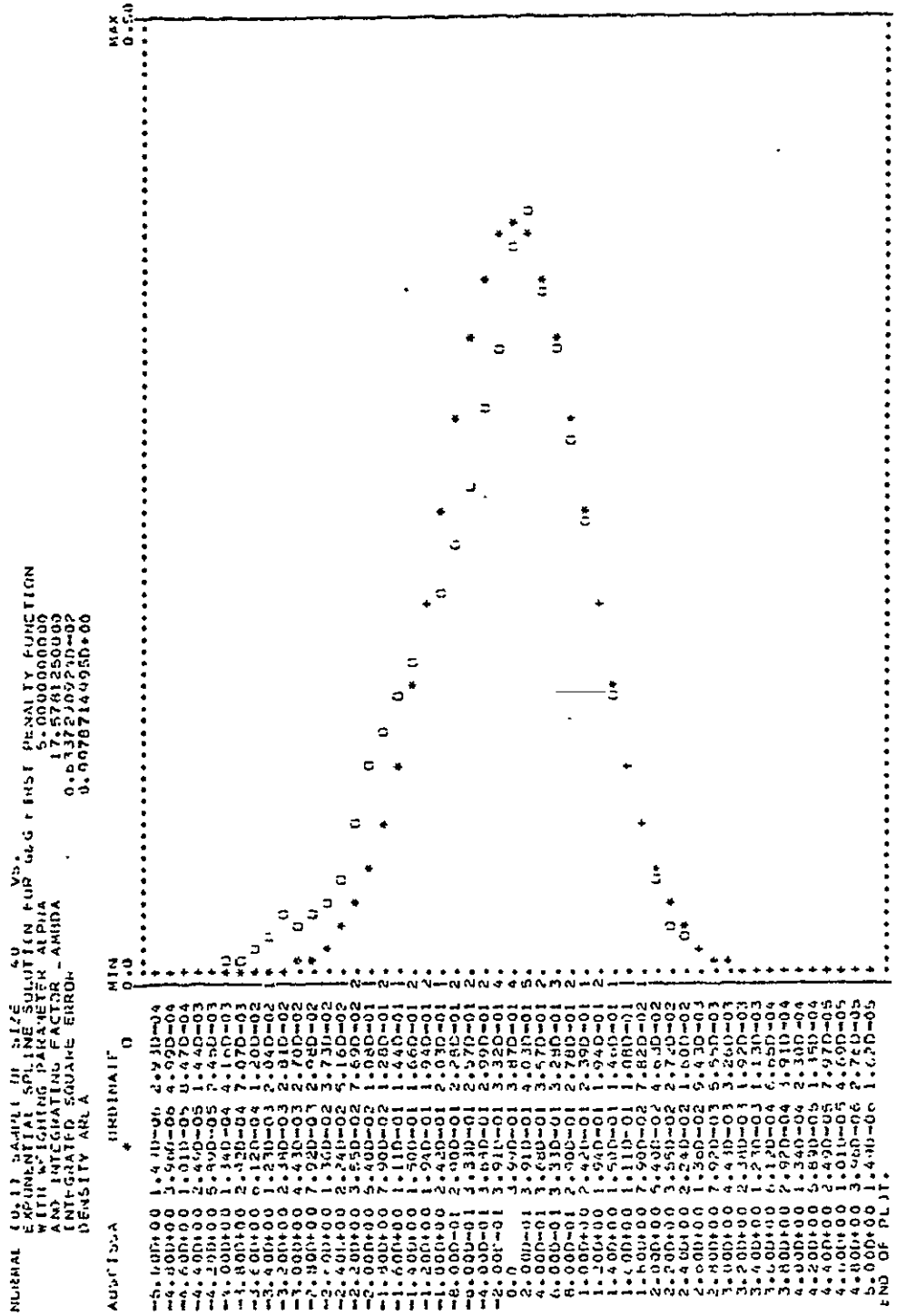


FIGURE 3.3.2 N = 40 N(0,1) Exponential Spline Solution



3.4 Some New Results

Consider the maximum penalized likelihood problem (3.2.2) with penalty functional

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} f'(x)^2 dx \ . \tag{3.4.1}$$

We choose the Sobolev space $H^1(-\infty,\infty)$ for the manifold $H(\Omega)$. As noted in section 3.2, in view of Lemma 3.2.1, Theorem 3.2.1 does not guarantee the existence of solutions to problem (3.2.2) with (3.4.1). We shall prove that the solution actually has finite support. From Theorem 3.2.2 we have that the solution is a monospline of degree 2 and that it is continuous. We begin by proving a useful inequality.

Lemma 3.2.1. Suppose $f \in H^1(-\infty,\infty)$ satisfies the constraints of problem (3.2.2). Then

$$f(x) \le \left(\frac{3}{2}\right)^{2/3} \left[\int_{-\infty}^{x} f(y)dy\right]^{1/3} \left[\int_{-\infty}^{x} f'(y)^2 dy\right]^{1/3} \ . \tag{3.4.2}$$

Proof. By the fundamental Theorem of Calculus we have

$$\begin{aligned} f^{3/2}(x) - f^{3/2}(a) &= \int_a^x \left[f^{3/2}(y)\right]' dy \\ &= \frac{3}{2} \int_a^x f^{1/2}(y) f'(y) dy \\ &\le \frac{3}{2} \left[\int_a^x f(y)dy\right]^{1/2} \left[\int_a^x f'(y)^2 dy\right]^{1/2} \end{aligned}$$

using the Cauchy-Schwarz inequality. Now let $a \to -\infty$. We have equality in (3.4.2) if and only if $f' = cf^{1/2}$. This proves the lemma.

Theorem 3.4.1. Consider the maximum penalized likelihood problem (3.2.2) for $f \in H^1(-\infty,\infty)$ and $\Phi(f)$ given by (3.4.1). Then we may restrict ourselves to functions supported on a finite interval (a,b) for a given sample $x_1,\ldots,x_N$.

Proof. We consider the properties of the maximum penalized likelihood criterion function

$$\hat{L}(f) = \sum_{i=1}^{N} \log f(x_i) - \alpha \int_{-\infty}^{\infty} f''(x)^2 dx \quad . \tag{3.4.3}$$

We assume without loss of generality that $-\infty < x_1 \leq x_2 \leq \ldots \leq x_N < \infty$. Since f integrates to one, we have from Lemma 3.4.1 the bound $\forall\, x \in (-\infty,\infty)$

$$f(x) \leq \left(\frac{3}{2}\right)^{2/3} \left[\int_{-\infty}^{\infty} f''(y)^2 dy\right]^{1/3} = C \quad . \tag{3.4.4}$$

Equations (3.4.4) and (3.4.3) imply

$$\hat{L}(f) \leq N \log C - \left(\frac{2}{3}\right)^2 C^3 \quad .$$

Thus as $C \to \infty$, $\hat{L}(f) \to -\infty$. Therefore, there exists $M < \infty$ such that the constraints of problem (3.2.2) are equivalent to

$$\begin{aligned} &\int_{-\infty}^{\infty} f(x)dx = 1 \\ &M \geq f \geq 0 \quad . \end{aligned} \tag{3.4.5}$$

Similarly,

$$\sum_{i=1}^{N} \log f(x_i) \leq \log f(x_k) + (N-1) \log M$$

for any $k = 1,N$. So as $f(x_k) \to 0$, $\hat{L}(f) \to -\infty$. Therefore, there exists $m > 0$ such that the constraints (3.4.5) are equivalent to

$$\begin{aligned} &\int_{-\infty}^{\infty} f(x)dx = 1 \\ &M \geq f \geq 0 \\ &f(x_k) \geq m \quad k = 1,N \quad . \end{aligned} \tag{3.4.6}$$

Now x_1 is the leftmost sample point. From Lemma 3.4.1 we have that

$$f(x_1) \leq \left(\frac{3}{2}\right)^{2/3} \left[\int_{-\infty}^{x_1} f(y)dy\right]^{1/3} \left[\int_{-\infty}^{x_1} f''(y)^2 dy\right]^{1/3} \tag{3.4.7}$$

with equality if and only if

$$f' = cf^{1/2} \quad \text{for} \quad x \in (-\infty, x_1] \quad . \tag{3.4.8}$$

Thus (3.4.8) implies the existence of $a > -\infty$ and $\gamma > 0$ such that

$$f(x) = \begin{cases} \gamma(x-a)^2 & \text{for} \quad x \in (a, x_1] \\ 0 & \text{for} \quad x \leq a \quad . \end{cases} \tag{3.4.9}$$

Notice that in (3.4.7) for all other things equal, $\int_{-\infty}^{x_1} f'(y)^2 dy$ is what we would like to minimize in order to maximize $\hat{L}(f)$.

We claim that we may restrict the constraint set (3.4.6) to the integral and nonnegativity constraints, along with

$$\{f : \exists a > -\infty \text{ with } f \text{ given by (3.4.9) for } -\infty < x \leq a\} \tag{3.4.10}$$

We wish to show that there exists $\hat{a}$ such that we may restrict the constraint set (3.4.10) to those f such that $a \geq \hat{a}$. If f is given by (3.4.9), then

$$\int_{-\infty}^{x_1} f(x)dx = \frac{1}{3} \gamma(x_1-a)^3$$

Thus

$$\gamma = 3(x_1-a)^{-3} \int_{-\infty}^{x_1} f(x)dx \tag{3.4.11}$$

Substituting (3.4.11) in (3.4.9) for $x = x_1$ implies

$$f(x_1) = 3(x_1-a)^{-1} \int_{-\infty}^{x_1} f(x)dx$$

$$\leq 3(x_1-a)^{-1} \quad . \tag{3.4.12}$$

Thus with m as in (3.4.6) we have

$$m \leq f(x_1) \leq 3(x_1-a)^{-1} \quad .$$

Solving for a ,

$$a \geq x_1 - 3/m \equiv \hat{a} \quad .$$

Since m depends on the data and x_1 is fixed, we have shown we may restrict ourselves to functions supported on $[\hat{a},\infty)$. We may use the same argument on the rightmost sample point x_N with the result

$$b \leq x_N + \frac{3}{m} \equiv \hat{b} \quad .$$

Thus we may restrict ourselves to functions supported on the finite interval $[\hat{a},\hat{b}]$, proving our theorem.

Corollary 3.4.1. The maximum penalized likelihood estimate considered in Theorem 3.4.1 is a monospline of degree 2.

Proof. From (3.4.9) we see that $f(a) = 0$ with the similar result $f(b) = 0$ following immediately. Thus we may restrict our manifold $H(\Omega)$ in problem (3.2.2) to $H_0^1(\hat{a},\hat{b})$. The corollary now follows immediately from Theorem 3.2.2.

Higher Derivatives

We conjecture that Theorem 3.4.1 generalizes to penalty functions of the form

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} f^{(s)}(x)^2 dx \qquad (3.4.13)$$

where $f \in H^{(s)}(-\infty,\infty)$. We shall present a weaker result for the case $s = 2$. We remark that our approach in the proof of Theorem 3.4.1 was the following: Suppose we are given any values of $f(x_k) > 0$ for $k = 1,N$. Then the likelihood portion of $\hat{L}(f)$ is fixed and we can only improve the penalty term. For any given positive area to the left of x_1 we see from (3.4.7) that there is a lower bound on the portion of the penalty functional (3.4.1) given by

$$\int_{-\infty}^{x_1} f'(x)^2 dx \quad .$$

This lower bound is attained only if f is the polynomial of degree 2 given by (3.4.9). For the case $s = 2$, a polynomial of degree 4 is the solution. However, since $f \in H^2(-\infty,\infty)$ implies that $f(x_1)$ and $f'(x_1)$

must be matched, we find that an arbitrary area to the left of x_1 cannot be attained.

Theorem 3.4.2. Consider the maximum penalized likelihood problem (3.2.2) with $f \in H^2(-\infty,\infty)$ and $\Phi(f)$ given by (3.4.13) with $s = 2$. Suppose the solution satisfies

$$\int_{-\infty}^{x_1} f(x)dx \leq \frac{3}{4}\frac{f^2(x_1)}{f'(x_1)} \quad \text{if} \quad f'(x_1) > 0$$

with no conditions required if $f'(x_1) \leq 0$, and similar conditions for the sample x_N. Then the solution is the monospline of degree 4 as in Theorem 3.2.1.

Proof. Consider

$$I(f) = \int_{-\infty}^{x_1} f''(x)^2 dx \tag{3.4.14}$$

which is related to the penalty functional. We claim that for any given values of

$$f(x_1),\ f'(x_1) \text{ and } \int_{-\infty}^{x_1} f(x)dx\ , \tag{3.4.15}$$

the optimal solution to problem (3.2.2) will minimize $I(f)$. This follows since we can only improve the penalty portion of $\hat{L}(f)$ by minimizing $I(f)$.

We now show that the solution to minimizing (3.4.14) given (3.4.15) is

$$f(x) = \begin{cases} a(x-x_o)^4 + b(x-x_o)^3 & x \in [x_o,x_1] \\ 0 & x \in (-\infty,x_o) \end{cases} \tag{3.4.16}$$

for some $-\infty < x_o < x_1$, $a \leq 0$, $b > 0$ picked to satisfy (3.4.15).

The second Gateaux derivative of $I(f)$ in the feasible directions ξ, η is

$$D^2I(f)(\eta,\xi) = 2\int_{-\infty}^{x_1} \xi''(x)\eta''(x)dx \ .$$

Since $D^2I(f)(\eta,\eta) > 0$, $I(f)$ is strictly convex so that (3.4.16) will be the unique solution. The tangent cone $T(f)$ (feasible directions) for η is defined by

$$\eta \in H_0^2(-\infty,x_1)$$
$$\int_{-\infty}^{x_1} \eta(x)dx = 0 \qquad (3.4.17)$$
$$\text{and if } f(x) = 0 \text{ , then } \eta(x) \geq 0 \ .$$

The necessary condition that f solve our problem is that

$$DI(f)(\eta) \geq 0 \quad \forall\ \eta \in T(f) \qquad (3.4.18)$$

or

$$f''(x_o)\eta'(x_o) - f'''(x_o)\eta(x_o) + \int_{x_o}^{x_1} f^{(iv)}(x)\eta(x)dx \geq 0 \qquad (3.4.19)$$

after integrating (3.4.18) by parts twice and noting that $\eta(x_1) = \eta'(x_1) = 0$. A fourth order polynomial satisfies (3.4.19). Since $f \in H^2$, the constant and linear terms vanish in the polynomial. By considering various $\eta \in T(f)$, we may show that the quadratic term vanishes, $b > 0$, $a \leq 0$. Let $L = x_1 - x_o$. Then

$$\int_{-\infty}^{x_1} f(x)dx = \frac{aL^5}{5} + \frac{bL^4}{4} \qquad (3.4.20)$$

where f is given by (3.4.16). Now a,b, and x_o are determined by the equations

$$f(x_1) = aL^4 + bL^3$$
$$f'(x_1) = 4aL^3 + 3bL^2$$
$$\int_{-\infty}^{x_1} f(x)dx = \frac{aL^5}{5} + \frac{bL^4}{4} \ .$$

It may be shown that if $f'(x_1) \leq 0$, then any area under the polynomial may be satisfied. However, if $f'(x_1) > 0$, then

$$\int_{-\infty}^{x_1} f(x)dx \leq \frac{3}{4} \frac{f^2(x_1)}{f'(x_1)} \quad .$$

If this constraint is not active, then we may proceed as in Theorem 3.4.1, proving our theorem.

IV. THE DISCRETIZED MAXIMUM PENALIZED LIKELIHOOD ESTIMATOR

4.1 Introduction

Since the infinite dimensional maximum penalized likelihood problem (3.2.2) for a general penalty function based on derivatives appears nontractable, we consider solving a finite dimensional problem motivated by the former. We deal with the nonnegativity constraint directly, thereby avoiding the unsatisfactory device of working with the square root of the density estimator. As our class of estimators, we shall consider simple functions and continuous piecewise linear functions with finite support defined, for convenience, by a uniform mesh on the interval (a,b). Specifically, we define the mesh by the nodes $a = t_0, t_1, \ldots, t_m = b$ where the mesh spacing h is given by $(b-a)/m$ and $t_{k+1} - t_k = h$ for all k. We define the k^{th} interval to be $I_k = [t_{k-1}, t_k)$.

Let $s(t)$ be a simple function defined on the mesh by

$$s(t) = s(t_k) \quad \forall\ t \in I_k \quad \text{for } k = 1,m \tag{4.1.1}$$

for m given values of $s(t_k)$ and zero elsewhere. Clearly,

$$\int_{-\infty}^{\infty} s(t)dt = \sum_{k=1}^{m} h\, s(t_k) \quad . \tag{4.1.2}$$

Similarly, let $p(t)$ be a continuous piecewise linear function defined on the mesh by

$$p(t) = p(t_{k-1}) + h^{-1}(t-t_{k-1})[p(t_k) - p(t_{k-1})] \quad t \in I_k \tag{4.1.3}$$

for $m+1$ given values of $p(t_k)$ and zero elsewhere. For p to be continuous, we define $p(t_0) = p(t_m) = 0$. It is easy to show that

$$\int_{-\infty}^{\infty} p(t)dt = \sum_{k=1}^{m-1} h\, p(t_k) \quad . \tag{4.1.4}$$

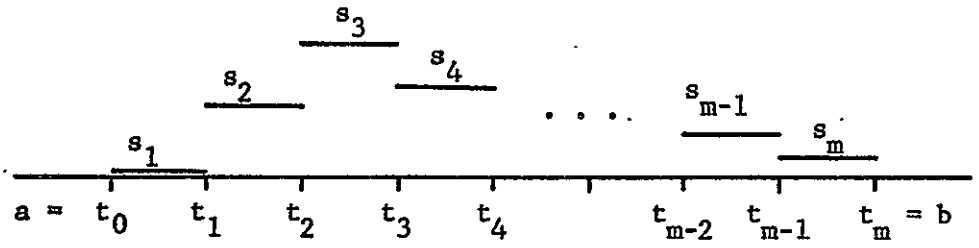
If we define

$$s_k = s(t_k) \quad k = 1,m$$

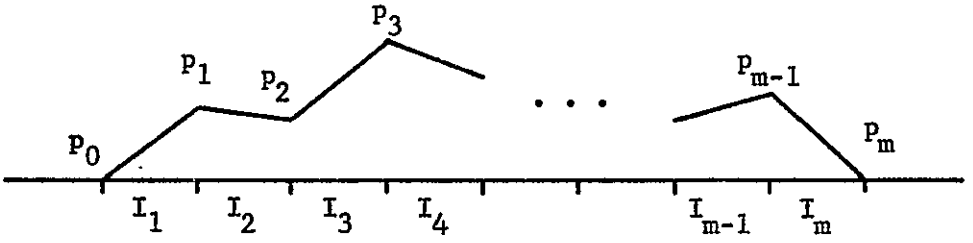
and

$$p_k = p(t_k) \quad k = 0,m$$

then typically, since $p_0 = p_m = 0$, we have



(4.1.5)



(4.1.6)

For the infinite dimensional problem, one criterion functional for a given sample set $x_1,\ldots,x_N$ was

$$\hat{L}(f) = \sum_{i=1}^{N} \log f(x_i) - \alpha \int_{-\infty}^{\infty} f'(t)^2 dt \quad . \tag{4.1.7}$$

We approximate the differential operator by finite differences over values at the mesh nodes. Similarly, we use the trapezoidal rule to approximate the integral operator. Thus we consider two criterion functions to be maximized that approximate (4.1.7)

$$L_s(s) = \sum_{i=1}^{N} \log s(x_i) - \alpha \sum_{k=1}^{m} h \left[\frac{s(t_k) - s(t_{k-1})}{h}\right]^2 \qquad (4.1.8)$$

and

$$L_p(p) = \sum_{i=1}^{N} \log p(x_i) - \alpha \sum_{k=1}^{m} h \left[\frac{p(t_k) - p(t_{k-1})}{h}\right]^2 . \qquad (4.1.9)$$

The subscript on L indicates whether simple or piecewise linear functions are under consideration. We may make approximations similar to (4.1.8) and (4.1.9) for higher order derivatives. To satisfy the definition of a density function, we place the following pairs of constraints on the functions $s(\cdot)$ and $p(\cdot)$, respectively:

$$\begin{aligned} & s(t_k) \geq 0 && p(t_k) \geq 0 \\ & \sum_{k=1}^{m} s(t_k) = \frac{1}{h} && \sum_{k=1}^{m-1} p(t_k) = \frac{1}{h} \end{aligned} \quad . \qquad (4.1.10)$$

With these constraints from (4.1.2) and (4.1.4), the functions will be non-negative on the real line and integrate to one.

We consider the problem of maximizing (4.1.8) or (4.1.9) for functions given by (4.1.1) or (4.1.3) under the constraints (4.1.10). We call solutions to these constrained optimization problems discretized maximum penalized likelihood estimates. In the next sections we consider the following: the existence and uniqueness of the discretized maximum penalized likelihood estimate, its consistency properties, and the sense in which the estimates approximate solutions to the corresponding infinite dimensional problem.

4.2 Existence and Uniqueness

Although we shall seldom retain more than one or two terms in the penalty functional, we may consider including r terms involving approximations of all derivatives up to the r^{th} order. Following our discussion in Section 3.1, pick a mesh $a = t_0, t_1, \ldots, t_m = b$ such that $t_{k+1} - t_k = h$. For convenience, extend this mesh to the entire real line.

Let $p(t)$ defined by (4.1.3) denote a continuous piecewise linear function that is identically zero outside the finite interval (a,b). Clearly, $p(\cdot)$ is determined by the $m-1$ values $p(t_k)$ for $k = 1,m-1$, since we define $p(t_0) = p(t_m) = 0$ for continuity. Consider the following constrained optimization problem for some fixed nonnegative weights α_j :

$$\text{maximize } L_p(p) = \sum_{i=1}^{N} \log p(x_i) - \sum_{j=1}^{r} \alpha_j \sum_{k=1}^{m+j-1} h\left[\frac{\nabla^j p(t_k)}{h^j}\right]^2 \quad (4.2.1)$$

subject to

$$p(t_k) \geq 0 \qquad k = 1, m-1$$
$$\sum_{k=1}^{m-1} p(t_k) = \frac{1}{h} \quad (4.2.2)$$

where

$$\nabla^j p(t_k) = (1-B)^j p(t_k)$$

where 1 and B are index shifting operators such that

$$1p(t_k) = p(t_k)$$

and

$$Bp(t_k) = p(t_{k-1}) \quad .$$

For example, $(1-B)^2 p(t_k) = p(t_k) - 2p(t_{k-1}) + p(t_{k-2})$, which would be the discrete approximation to h^2 times the second derivative. Again, the constraints (4.2.2) guarantee that $p(\cdot)$ will be nonnegative and integrate to one. Although the following are really corollaries to Proposition 3.2.1, we give below a simple finite dimensional proof not possible for use in the more general infinite dimensional case.

Theorem 4.2.1. Suppose we are given a sample $x_1,\ldots,x_N$, each contained in the finite interval (a,b) partitioned in an equally spaced mesh $t_0,\ldots,t_m$. Consider the problem of maximizing (4.2.1) over all continuous piecewise linear functions given by (4.1.3) under the constraints (4.2.2). Then solutions to this problem exist and are unique.

Proof. The constraint set (4.2.2) is clearly a convex and compact set in R^{m-1}. Since $L_p(\cdot)$ is a continuous functional in the values $p(t_k)$, we have existence of a global maximizer.

To prove uniqueness, we suppose that we have two continuous piecewise linear functions p_1 and p_2 such that $L_p(p_1) = L_p(p_2)$. Consider the continuous piecewise linear function defined by

$$p(t) = \tfrac{1}{2}p_1(t) + \tfrac{1}{2}p_2(t) \quad . \tag{4.2.3}$$

This function is admissible since it satisfies the constraints (4.2.2). Since the log is a strictly concave function, we have

$$\sum_{i=1}^{N} \log p(x_i) > \tfrac{1}{2}\sum_{i=1}^{N} \log p_1(x_i) + \tfrac{1}{2}\sum_{i=1}^{N} \log p_2(x_i) \quad . \tag{4.2.4}$$

Also we have by linearity that

$$\nabla^j p(t_k) = \tfrac{1}{2}\nabla^j p_1(t_k) + \tfrac{1}{2}\nabla^j p_2(t_i) \tag{4.2.5}$$

For any two real numbers a and b

$$(a + b)^2 \leq 2a^2 + 2b^2 \quad . \tag{4.2.6}$$

Combining (4.2.5) and (4.2.6) and summing implies

$$\sum_k [\nabla^j p(t_k)]^2 \leq \tfrac{1}{2}\sum_k [\nabla^j p_1(t_k)]^2 + \tfrac{1}{2}\sum_k [\nabla^j p_2(t_k)]^2 \quad . \tag{4.2.7}$$

After multiplying (4.2.7) by $\alpha_j h^{1-j}$ and summing over j together with (4.2.4), we have

$$L_p(p) > \tfrac{1}{2}L(p_1) + \tfrac{1}{2}L(p_2) = L(p_1)$$

Thus the existence of two distinct global maximizers would lead to a contradiction.

Theorem 4.2.2. Under the conditions of Theorem 4.2.1, consider the problem of maximizing $L_s(\cdot)$ corresponding to (4.2.1) over all simple functions given by (4.1.1) under the constraints (4.1.10). Then solutions to this

problem exist and are unique.

Proof. The constraint region (4.1.10) is convex and compact. The objective functional $L_s(\cdot)$ is a continuous function of the values $s(t_k)$. This proves existence of solutions. The uniqueness follows exactly as in Theorem 4.2.1.

4.3 Consistency of the Discretized Maximum Penalized Likelihood Estimator

In this section we prove that the simple function maximum penalized likelihood estimator is consistent in mean square error. For large sample sizes, this estimator looks very much like the histogram. In later section we consider how the continuous piecewise linear function and the simple function are "close" for the same data.

Consider the simple function $s(\cdot)$ discussed in Section 4.1. Again, for convenience, we define $s_k = s(t_k)$ for $k = 1,m$. Extending the mesh over the entire real line, we have $s_{-1} = s_{m+3} = 0$, for example. For a given sample $x_1,\ldots,x_N$, let

$$\begin{aligned} \nu_0 &= \#\text{ samples in } (-\infty,a) \\ \nu_k &= \#\text{ samples in } I_k = [t_{k-1},t_k) \quad k = 1,m \\ \nu_{m+1} &= \#\text{ samples in } [b,\infty) \end{aligned} \tag{4.3.1}$$

Then $\sum_{k=0}^{m+1} \nu_k = N$ and we define

$$n' = \sum_{k=1}^{m} \nu_k \quad . \tag{4.3.2}$$

We consider truncating our data outside the interval $[a,b)$. This will introduce a slight bias into the solution. Then using (4.3.1), our objective functional with the first derivative penalty functional becomes

$$\text{maximize } L_s(s) = \sum_{k=1}^{m} \nu_k \log s_k - \frac{\alpha}{h} \sum_{k=0}^{m} (s_{k+1}-s_k)^2 \tag{4.3.3}$$

$$\text{subject to } s_k \geq 0 \quad k = 1,m \tag{4.3.4}$$

$$\sum_{k=1}^{m} s_k = \frac{1}{h} \ .$$

We may now prove consistency of solutions to (4.3.3).

Theorem 4.3.1. Let $x_1,\ldots,x_N$ be a random sample from a continuous density f_o. Then the simple discretized maximum penalized likelihood estimator solving (4.3.3) is asymptotically consistent in the mean square error in the following sense: if we pick $h = h(N)$ so that as $N \to \infty$ and $h(N) \to 0$ we have $Nh(N)^2 \to \infty$, then we may make the mean square error arbitrarily small by taking the interval (a,b) sufficiently large.

Proof. The constraint $s_k \geq 0$ is not active for those k where $\nu_k > 0$, since $s_k^* = 0$ with $\nu_k > 0$ implies that $L_s(s^*) = -\infty$. However, for the feasible choice

$$\hat{s}_k = \frac{\nu_k}{n'h} \quad k = 1,m$$

we may calculate

$$L_s(\hat{s}) = \sum_k \nu_k \log\left(\frac{\nu_k}{n'h}\right) - \frac{\alpha}{h} \sum_k \left(\frac{\nu_{k+1}}{n'h} - \frac{\nu_k}{n'h}\right)^2$$

$$\geq \sum_k \nu_k \log \nu_k - \sum_k \nu_k \log n'h - \frac{\alpha}{n'^2h^3} n'^2$$

$$\geq -n' \log n'h - \frac{\alpha}{h^3} > -\infty \ .$$

Ignoring the inequality constraint, we have from the theory of Lagrange multipliers that there exists $\lambda \in R$ such that

$$\frac{\partial L_s}{\partial s_i} + \lambda \frac{\partial}{\partial s_i}\left[\sum_{k=1}^{m} s_k - \frac{1}{h}\right] = 0 \quad i = 1,m \quad . \tag{4.3.5}$$

For our problem, equation (4.3.5) becomes

$$\frac{\nu_i}{s_i} + \frac{2\alpha}{h} \nabla^2 s_{i+1} + \lambda = 0 \quad i = 1,m \tag{4.3.6}$$

where

$$\nabla^2 s_{i+1} = s_{i+1} - 2s_i + s_{i-1} \quad . \tag{4.3.7}$$

Multiply (4.3.6) by s_i and sum over $i = 1,m$. Using (4.3.2) and the second constraint in (4.3.4), we may solve for the Lagrange multiplier as

$$\lambda = n'h - 2\alpha \sum_{i=1}^{m} s_i \nabla^2 s_{i+1}$$

Substituting this value in (4.3.6), our necessary conditions become, after dividing by N ,

$$\frac{\nu_i}{Ns_i} + \frac{2\alpha}{Nh} \nabla^2 s_{i+1} - \frac{n'}{N} h - \frac{2\alpha}{N} \sum_{j=1}^{m} s_j \nabla^2 s_{j+1} = 0 \tag{4.3.8}$$

Before solving for s_i , let us bound the second and fourth terms in (4.3.8). From the constraints (4.3.4) we see that

$$0 \le s_i \le \frac{1}{h} \quad i = 1,m \quad . \tag{4.3.9}$$

Using (4.3.9) and the definition in (4.3.7), we arrive at the following nonstochastic bounds:

$$\frac{-2}{h} \le \nabla^2 s_i \le \frac{1}{h} \quad i = 1,m \tag{4.3.10}$$

Since $s(t)$ is zero outside (a,b) , we have

$$\begin{aligned} \sum_{j=1}^{m} s_j \nabla^2 s_{j+1} &= \sum_{j=1}^{m} s_j (s_{j+1} - 2s_j + s_{j-1}) \\ &= \sum_{j=1}^{m} 2s_j s_{j+1} - 2s_j^2 \\ &= -s_1^2 - s_m^2 - \sum_{j=1}^{m-1} (s_{j+1} - s_j)^2 \quad . \end{aligned} \tag{4.3.11}$$

Using (4.3.11), we arrive at the bounds

$$-\frac{2}{h^2} \leq \sum_{j=1}^{m} s_j \nabla^2 s_{j+1} \leq 0 \quad . \tag{4.3.12}$$

Using the bounds (4.3.10) and (4.3.12) in equation (4.3.8), the necessary conditions become

$$\frac{\nu_i}{N s_i} + \frac{2\alpha}{Nh} O(\frac{1}{h}) - \frac{n'}{N} h - \frac{2\alpha}{N} O(\frac{1}{h^2}) = 0 \quad . \tag{4.3.13}$$

Suppose f_o is the true sampling density. We define

$$\epsilon = 1 - \int_a^b f_o(x)dx \tag{4.3.14}$$

and

$$p_k = \int_{I_k} f_o(x)dx \quad \text{for } k = 1,m \quad . \tag{4.3.15}$$

Now as $N \to \infty$ keeping h fixed

$$\frac{\nu_k}{N} \to p_k \quad \text{in quadratic mean} \tag{4.3.16}$$

and

$$\frac{n'}{N} \to 1 - \epsilon \quad \text{in quadratic mean} \tag{4.3.17}$$

since the variances of the quantities in (4.3.16) and (4.3.17) vanish as $N \to \infty$. Thus as $N \to \infty$ keeping h fixed, we have from (4.3.13)-(4.3.17)

$$s_i \to \frac{p_i}{h} \frac{1}{1-\epsilon} \quad \text{in quadratic mean} \quad . \tag{4.3.18}$$

Since f_o is continuous, we have that as $h \to 0$

$$\frac{p_i}{h} \to f_o(x_i) \tag{4.3.19}$$

where x_i is the point in I_i as $h \to 0$. Therefore, if we demand as $N \to \infty$ and $h \to 0$ that $Nh^2 \to \infty$, we see from (4.3.13), (4.3.18), and (4.3.19)

$$s_i \to \frac{f_o(x_i)}{1-\epsilon} \quad \text{in quadratic mean} . \tag{4.3.20}$$

Thus the mean square error of the discretized maximum penalized likelihood estimate at a point x_i is seen to be

$$\text{m.s.e.} \to f_o(x_i)^2 \left(\frac{\epsilon}{1-\epsilon}\right)^2 . \tag{4.3.21}$$

By picking the interval (a,b) arbitrarily large, we may choose ϵ arbitrarily small in (4.3.21). This proves our theorem.

Consider generalizing the objective functional (4.3.3) to include the r^{th} derivative in the penalty term. Using the notation of equation (4.2.1) our problem becomes

$$\text{maximize} \quad L_s(s) = \sum_{k=1}^{m} \nu_k \log s_k - \frac{\alpha}{h^{2r-1}} \sum_k [\nabla^r s(t_k)]^2 \tag{4.3.22}$$

subject to the constraints (4.3.4). The analogous result to Theorem 4.3.1 is

Theorem 4.3.2. The simple discretized maximum likelihood estimator solving (4.3.22) is consistent in the sense given in Theorem 4.3.1 if we pick $h(N)$ so that as $N \to \infty$ and $h(N) \to 0$, we have $Nh(N)^{2r} \to \infty$.

Proof. The proof is parallel to that of Theorem 4.3.1.

The Truncated Density

The effect of throwing away data points outside the interval (a,b) when solving problem (4.3.3) is that we are really estimating the density

$$g(x) = \begin{cases} \dfrac{f_o(x)}{1-\epsilon} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{4.3.23}$$

where we call $g(x)$ the truncated density and ϵ is defined in (4.3.14).

Nonparametric estimates in the tails are generally unacceptable as we discussed in section 2.2. However, in situations where nonparametric density estimation is appropriate, faithful representation of the unknown density near the modes and in area of high density is the issue. This is the goal of the discretized maximum penalized likelihood estimator.

Corollary 4.3.1. Under the conditions of Theorem 4.3.1 for a fixed interval (a,b), the discretized maximum penalized likelihood estimate is consistent with the truncated density (4.3.23).

Proof. Follows immediately from the proof of Theorem 4.3.1 and equation (4.3.20).

If we assume that the sampling density is absolutely continuous, we have the following consistency result:

Theorem 4.3.3. Suppose the sampling density $f_o(x)$ is absolutely continuous. Let $g(x)$ denote the truncated density (4.3.23) for some interval $(-A,A)$. Then the simple discrete penalized estimator $s_N(x)$ (where N denotes the sample) is consistent on $(-A,A)$ in the integrated mean square error; that is,

$$\lim_{N\to\infty} \int_{|x|<A} E|s_N(x) - g(x)|^2 dx = 0$$

where s_N solves (4.3.22).

Proof. Consider the finite interval $(-A,A)$; divide it into m intervals of equal lengths $h > 0$; that is

$$m = \frac{2A}{h}$$

$$I_1 \cup I_2 \cup \ldots \cup I_m = (-A,A) \quad .$$

Let x_k be a fixed point in I_k for $k = 1,m$. From Corollary 4.3.1, we have that $s_N(x_k)$ converges in mean square to $g(x_k)$. Therefore, given $\epsilon > 0$, there exists $h > 0$ sufficiently small and $n_{x_k} < \infty$ such that

$$E|s_N(x_k) - g(x_k)|^2 < \epsilon \quad \forall\, N \geq n_{x_k} \quad \text{for} \quad k = 1,m \tag{4.3.24}$$

where N is also picked large enough so that $Nh^{2r} \to \infty$ and $h \to 0$.

Since there are a finite number of intervals m , we define

$$n^*_\epsilon = \max_{1\leq k\leq m} n_{x_k} < \infty .$$

Thus (4.3.24) holds for all k .

For a given h consider

$$\delta_h = \max_{1\leq k\leq m} \sup_{x,y\in I_k} |g(x) - g(y)| . \tag{4.3.25}$$

Now f_o and hence g are absolutely continuous on $(-A,A)$. Therefore, by absolute continuity

$$\lim_{h\to 0} \delta_h = 0 \tag{4.3.26}$$

Consider the mean square error at an arbitrary point $y \in I_k$. We have since $s_N(y) = s_N(x_k)$

$$\begin{aligned} E|s_N(y) - g(y)|^2 &= E|s_N(x_k) - g(y)|^2 \\ &\leq E|s_N(x_k) - g(x_k)|^2 + |g(x_k) - g(y)|^2 \\ &\leq \epsilon + \delta_h^2 \qquad \forall\, N \geq n^* \end{aligned} \tag{4.3.27}$$

under the conditions (4.3.24) using the triangle inequality and (4.3.25). Since (4.3.27) holds for any y we have

$$\int_{|x|<A} E|s_N(y) - g(y)|^2 dy \leq 2(\epsilon + \delta_h^2)A . \tag{4.3.28}$$

Now ϵ is arbitrarily small, $\delta_h^2 \to 0$ by (4.3.26), and A is fixed. This proves the theorem.

4.4 Approximation Results

Theorem 4.4.1. Suppose (a,b) is a finite interval and $x_1,\ldots,x_N$ a fixed sample of size N. For the penalty functional

$$\Phi(f) = \alpha\int_a^b f'(t)^2dt$$

we consider the discrete and infinite dimensional maximum penalized likelihood problems, truncating data if necessary. Then the simple function solution approaches the $H_0^1(a,b)$ monospline in $L^2(a,b)$ as $h \to 0$.

Proof. We denote the simple function solution by $s_h(\cdot)$ to emphasize the mesh spacing. Let $s_h(\cdot)$ be defined as in (4.4.1) and (4.4.5). For convenience let $s_h(t_1) = s_h(t_m) = 0$ (see 4.4.10). The two criterion functions are

$$L_h(s_h) = \sum_{i=1}^N \log s_h(x_i) - \frac{\alpha}{h}\sum_k [s_h(t_k) - s_h(t_{k-1})]^2 \qquad (4.4.1)$$

and

$$L_f(f) = \sum_{i=1}^N \log f(x_i) - \alpha\int_a^b f'(t)^2dt \qquad (4.4.2)$$

Step 1. Let f^* denote the solution to the continuous problem (4.4.2). By Theorem 3.2.2 we know f^* exists uniquely and is a monospline of degree two. Let s_h^* be the unique solution to the discrete problem (4.4.1) for a given h.

Claim. We can find $s_{f^*,h}$ a simple function approximation to f^* that satisfies the discrete problem constraints such that

$$L_h(s_{f^*,h}) \to L_f(f^*) \qquad (4.4.3)$$

in the sup norm as $h \to 0$.

Proof of Claim. We construct $s_{f^*,h}$ and demonstrate the desired properties. Let

$$s_{f^*,h}(t_k) = \frac{1}{h}\int_{I_k} f^*(t)dt \qquad k = 1,m \quad . \tag{4.4.4}$$

Then $s_{f^*,h}$ is nonnegative and integrates to one. Since f^* is a monospline of degree two, f^* is infinitely differentiable except perhaps at the sample points x_i and at two points between adjacent samples and at one point between an interval endpoint and the extreme sample. Thus there are nc more than $3N$ points of derivative discontinuities in all. For h small enough, no interval I_k will contain more than one such point of discontinuity. Where they exist, all derivatives of f^* are bounded.

For $x \in I_k$ and some $a \in I_k$

$$s_{f^*,h}(x) = f^*(a) \quad . \tag{4.4.5}$$

Since

$$f^*(x) - f^*(a) = \int_a^x f^{*\prime}(y)dy$$

(4.4.5) and the Candy-Schwartz inequality imply

$$|f^*(x) - f(a)|^2 \le \int_a^x f^{*\prime}(y)^2dy \cdot \int_a^x dy$$

or

$$|f^*(x) - s_{f^*,h}(x)| \le \sqrt{h}\left[\int_{I_k} f^{*\prime}(y)^2dy\right]^{\frac{1}{2}} \quad . \tag{4.4.6}$$

Therefore, $s_{f^*,h}(x) \to f^*(x)$ in the sup norm as $h \to 0$ so that the log likelihood terms in (4.4.1) and (4.4.2) agree as $h \to 0$. We now consider the penalty terms. Let $s_{k,h} \equiv s_{f^*,h}(t_k)$, $k = 1,m$. Then using the Mean Value Theorem

$$\frac{1}{h}\sum_k (s_{k,h} - s_{k-1,h})^2 = \frac{1}{h}\sum_k \left(\frac{1}{h}\int_{I_k} f^*(t)dt - \frac{1}{h}\int_{I_{k-1}} f^*(t)dt\right)^2$$

$$= \frac{1}{h}\sum_k \left(f^*(\hat{x}_k) - f^*(\hat{x}_{k-1})\right)^2 \tag{4.4.7}$$

where $\hat{x}_k$ is a point in I_k. Letting $\tilde{x}_k$ denote the midpoint of I_k,

we can take a Taylor expansion and calculate (4.4.7) to be

$$\frac{1}{h}\sum_k (f^*(\tilde{x}_k) - f^*(\tilde{x}_{k-1}) + d_k\alpha_k h - d_{k-1}\alpha_{k-1}h)^2$$

where $d_k = \frac{d}{dt} f^*(t)$ evaluated at some point in I_k and $\tilde{x}_k - \hat{x} = \alpha_k h$ for $|\alpha_k| \le \frac{1}{2}$. Squaring in the above we obtain

$$\frac{1}{h}\sum_k (f^*(\tilde{x}_k) - f^*(\tilde{x}_{k-1}))^2 + \frac{1}{h}\sum_k h^2(d_k\alpha_k - d_{k-1}\alpha_{k-1})^2$$

$$+ \frac{2}{h}\sum_k [f^*(\tilde{x}_k) - f^*(\tilde{x}_{k-1})][h(d_k\alpha_k - d_{k-1}\alpha_{k-1})] \quad . \qquad (4.4.8)$$

Ignoring the finite number of points where $f^{*\prime}$ is discontinuous, $f^*(\tilde{x}_k) - f^*(\tilde{x}_{k-1})$ and $d_k - d_{k-1}$ are $0(h)$. The summing process in $0(m) = 0(\frac{1}{h})$ so that the second term is $\frac{1}{h}[0(h)]^2 = 0(h)$. Likewise the third term is $\frac{1}{h}\, 0(\frac{1}{h})0(h)h0(h) = 0(h)$. So it suffices to show that as $h \to 0$, the first term in (4.4.8) approximates $\int f^{*\prime}(t)^2 dt$ as $h \to 0$. Using the Fundamental Theorem of Calculus, we have that the difference of the first term in (4.4.8) and the continuous penalty term is

$$\frac{1}{h}\sum_k (f^*(\tilde{x}_k) - f^*(\tilde{x}_{k-1}))^2 - \int f^{*\prime}(t)^2 dt$$

$$= \frac{1}{h}\sum_k \Big[\int_{\tilde{x}_{k-1}}^{\tilde{x}_k} f^{*\prime}(t)dt\Big]^2 - \sum_k \int_{\tilde{x}_{k-1}}^{\tilde{x}_k} f^{*\prime}(t)^2 dt$$

$$= \frac{1}{h}\sum_k [f^{*\prime}(\xi_k)(\tilde{x}_k - \tilde{x}_{k-1})]^2 - \sum_k f^{*\prime}(\eta_k)^2(\tilde{x}_k - \tilde{x}_{k-1})$$

$$= \sum_k h[f^{*\prime}(\xi_k)^2 - f^{*\prime}(\eta_k)^2] \qquad (4.4.9)$$

since $\tilde{x}_k - \tilde{x}_{k-1} = h$ where ξ_k and $\eta_k \in I_k$. Now the term in brackets in (4.4.9) is

$$[f^{*\prime}(\xi_k) - f^{*\prime}(\eta_k)][f^{*\prime}(\xi_k) + f^{*\prime}(\eta_k)] \quad .$$

The first factor is $0(h)$ and the second term is bounded. Since the summing

process is $0(\frac{1}{h})$, (4.4.9) is $0(\frac{1}{h})h0(h) \to 0$ as $h \to 0$. In the finite number of intervals containing the derivative discontinuities, the contributions to the sum are less than $3N[0(h)] \to 0$, so that we could ignore those intervals in the above arguments. This proves the claim.

Step 2. Recall that s_h^* is the unique maximizer of (4.4.1). We have that $L_h(s_h^*) \geq L_h(s_{f^*,h}) > -\infty$ since $L_h(s_{f^*,h}) \sim L_f(f^*) > -\infty$. Therefore

$$\sup_{h \to 0} |s_h^*(t_k) - s_h^*(t_{k-1})| \to 0 \tag{4.4.10}$$

since otherwise the penalty term would tend to $-\infty$. We use (4.4.10) to approximate s_h^* with an H_0^1 function. Given s_h^* we define a continuous approximation f_h^* to s_h^* in the following way: let f_h^* be the piecewise linear function connecting the simple function s_h^* at the midpoints of the intervals. With this choice, f_h^* is nonnegative and integrates to one. If we consider the derivative approximation from the midpoint of I_{k-1} to I_k , it is d^2/h for the simple function, where $d = s_h^*(t_k) - s_h^*(t_{k-1})$, and for the piecewise linear function

$$\int_h f_h^{*\prime}(t)^2 dt = \int_h (\frac{d}{h})^2 dt = \frac{d^2}{h} \quad .$$

Therefore, by construction we have

$$\int f_h^{*\prime}(t)^2 dt = \frac{1}{h} \sum_k [s_h^*(t_k) - s_h^*(t_{k-1})]^2 \tag{4.4.11}$$

so that the penalty terms agree for any h . From (4.4.10) we see that f_h^* converges to s_h^* pointwise in the sup norm by the construction of f_h^* . Combining this fact with (4.4.11), we have

$$\|L_h(s_h^*) - L_f(f_h^*)\|_{L^\infty} \to 0 \quad \text{as} \quad h \to 0 \; . \tag{4.4.12}$$

Since both f_h^* and s_h^* are density functions, we have $\|f_h^* - s_h^*\|_{L^1} < 2$.

This bound, (4.4.12) and Holder's inequality imply

$$\|f_h^* - s_h^*\|_{L^2} \to 0 \quad \text{as} \quad h \to 0 \ . \tag{4.4.13}$$

Step 3. By their respective optimality properties

$$L_f(f_h^*) \le L_f(f^*)$$

and (4.4.14)

$$L_h(s_{f^*,h}) \le L_h(s_h^*) \ .$$

Combining (4.4.14) and (4.4.12), we have

$$L_h(s_{f^*,h}) \le L_h(s_h^*) \overset{h\to 0}{\to} L_f(f_h^*) \le L_f(f^*) \ . \tag{4.4.15}$$

But since $L_h(s_{f^*,h}) \to L_f(f^*)$ in the L^∞-norm as $h \to 0$ by (4.4.3), we have from (4.4.15)

$$L_f(f_h^*) \to L_f(f^*) \quad \text{as} \quad h \to 0 \ . \tag{4.4.16}$$

By the uniform strict concavity of $L_f(\cdot)$ with respect to the H^1 norm we have from (4.4.16)

$$\|f_h^* - f^*\|_{H^1} \to 0 \quad \text{as} \quad h \to 0 \ . \tag{4.4.17}$$

Now H^1 convergence implies L^2 convergence. Both f_h^* and s_h^* are in L^2. Therefore, the triangle inequality using (4.4.13) and (4.4.17) implies

$$\|s_h^* - f^*\|_{L^2} \to 0 \quad \text{as} \quad h \to 0 \ .$$

This proves the theorem.

Theorem 4.4.2. Under the same conditions as Theorem 4.4.1, the continuous piecewise linear solution approaches the $H_0^1(a,b)$ monospline in H^1 as $h \to 0$.

Proof. We replace equation (4.4.1) with

$$L_h(p_h) = \sum_{i=1}^{N} \log p_h(x_i) - \frac{\alpha}{h} \sum_k [p_h(t_k) - p_h(t_{k-1})]^2 \tag{4.4.18}$$

where p_h is a continuous piecewise linear function defined on the mesh of interval width h.

Step 1. We can find $p_{f^*,h}$ a continuous piecewise linear function approximation to f^* that satisfies the discrete problem constraints such that

$$L_h(p_{f^*,h}) \to L_f(f^*) \tag{4.4.19}$$

in the sup norm as $h \to 0$. Consider the function f^*_{2h} which approximated s^*_{2h} in step 2 of the proof of Theorem 4.4.1. f^*_{2h} has mesh nodes exactly at t_k since it has its nodes at the midpoints of the mesh with interval width $2h$. If we let $p_{f^*,h} = f^*_{2h}$, then by (4.4.17) $p_{f^*,h}$ is H^1 convergent to f^* and (4.4.19) is seen to hold by construction.

Step 2. Let p^*_h denote the maximizer of (4.4.18). We have $L_h(p^*_h) > L_h(p_{f^*,h}) > -\infty$ since $L_h(p_{f^*,h}) \sim L_f(f^*) > -\infty$. Thus (4.4.10) holds for p^*_h and $p^*_h \in H^1_0(a,b)$ as $h \to 0$. Therefore,

$$\|L_h(p^*_h) - L_f(p^*_h)\|_{L^\infty} \to 0 \quad \text{as } h \to 0 . \tag{4.4.20}$$

Step 3. By their respective optimality properties and (4.4.20)

$$L_h(p_{f^*,h}) \le L_h(p^*_h) \overset{h\to 0}{\to} L_f(p^*_h) \le L_f(f^*) \quad . \tag{4.4.21}$$

Using the strict concavity of $L_f(\cdot)$, (4.4.19), and the triangle inequality we may show exactly as in Theorem 4.4.1

$$\|p^*_h - f^*\|_{H^1} \to 0 \quad \text{as} \quad h \to 0$$

proving the theorem.

Plausible Theorem 4.4.3. Under the same conditions as Theorem 4.4.1 with the penalty functional

$$\Phi(f) = \alpha \int_a^b f''(t)^2 dt$$

the continuous piecewise linear solution approaches the $H_0^2(a,b)$ monospline in H^1 as $h \to 0$.

Remark. The H^1 convergence is the best possible since the discrete solution is not in H^2.

V. NUMERICAL IMPLEMENTATION AND SIMULATION RESULTS

5.1 The Numerical Algorithm

In presenting the numerical solution for the discrete maximum likelihood penalized problem, we choose the continuous piecewise linear solution rather than the simple function solution for its smoothness and approximation properties. We consider the penalty functional based on second differences which may be generalized to an arbitrary derivative approximation.

Let $t_1,\ldots,t_m$ be a given fixed mesh with mesh interval $h = t_{k+1}-t_k$, for $k = 1,m-1$. The continuous piecewise linear solution is defined as $p(t)$ and is determined by the values at the nodes which we denote

$$p_k = p(t_k) \quad k = 1,m$$

where (5.1.1)

$$p_1 = p_2 = p_{m-1} = p_m = 0$$

for convenience. The solution may be evaluated at a point t by

$$p(t) = \begin{cases} p_k + \dfrac{p_{k+1} - p_k}{h}(t - t_k) & t \in [t_k, t_{k+1}) \\ 0 & t \notin (t_2, t_{m-1}) \end{cases} \tag{5.1.2}$$

Let $x_1,\ldots,x_N$, be a random sample. We truncate those points not falling in the interval (t_2,t_{m-1}) and label the remaining points $x_1,\ldots,x_N$ To evaluate $p(\cdot)$ at x_i, we introduce the star indexing function $*:I \to I$ defined by

$$x_i \in [t_{*(i)}, t_{*(i)+1}) \quad i = 1,N \tag{5.1.3}$$

Thus the star function points to the interval in which x_i falls. Our criterion function (4.1.9) may be written as

$$\text{minimize } \frac{\alpha}{h^3} \sum_{k=2}^{m-1} (p_{k-1} - 2p_k + p_{k+1})^2$$

$$- \sum_{i=1}^{N} \log\left[p_{*(i)} + \frac{p_{*(i)+1} - p_{*(i)}}{h} (x_i - t_{*(i)})\right] \qquad (5.1.4)$$

subject to

$$p_k \geq 0 \quad k = 3,m-2$$

$$\sum_{k=3}^{m-2} p_k = \frac{1}{h} \qquad (5.1.5)$$

We may deal with the nonnegativity constraint in (5.1.5) directly by the substitution

$$\omega_k^2 = p_k \quad k = 1,m \qquad (5.1.6)$$

and solving for $\omega_3,\ldots,\omega_{m-2}$. From the theory of Lagrange multipliers there exists $\lambda \in R$ such that

$$4\omega_k \frac{\alpha}{h^3} \nabla^4 \omega_{k+2}^2 - 2\omega_k \sum_{*(i)=k} \frac{1 - \frac{x_i - t_k}{h}}{\omega_k^2 + \frac{\omega_{k+1}^2 - \omega_k^2}{h} (x_i - t_k)}$$

$$- 2\omega_k \sum_{*(i)=k-1} \frac{\frac{x_i - t_{k-1}}{h}}{\omega_{k-1}^2 + \frac{\omega_k^2 - \omega_{k-1}^2}{h} (x_i - t_{k-1})} - 2\lambda \qquad (5.1.7)$$

is identically zero for $k = 3,\ldots,m-2$ at the solution of problem (5.1.4) where

$$\nabla^4 p_{k+2} = p_{k+2} - 4p_{k+1} + 6p_k - 4p_{k-1} + p_{k-2}$$

Equations (5.1.7) along with the integral constraint

$$\frac{1}{h} - \sum_{k=3}^{m-2} \omega_k^2 = 0 \qquad (5.1.8)$$

determine $m - 3$ nonlinear equations in the $m - 3$ unknowns $\lambda,\omega_3,\ldots,\omega_{m-2}$.

Given initial (nonzero) estimates for the parameters, Newton's method is used to find the zeroes of the equations (5.1.7) and (5.1.8). This algorithm involves calculating the (m-3)x(m-3) symmetric Jacobian matrix and solving the resulting linear system for the changes $\Delta\omega_k$ and $\Delta\lambda$ in the estimates of the last iteration. Iteration is stopped when

$$[(\Delta\lambda)^2 + \sum_{k=3}^{m-2} (\Delta\omega_k)^2]^{\frac{1}{2}} \leq 10^{-5} \qquad (5.1.9)$$

Then $p(t)$ is determined by (5.1.6) and (5.1.2).

We emphasize that the number of mesh nodes m determines the amount of work necessary in the numerical solution of the discretized maximum penalized likelihood estimate (D.M.P.L.E.). The sample size N is important in calculating the Jacobian matrix, but only in a linear fashion. Thus the major effort of solving the (m-3)x(m-3) linear system does not depend on the sample size.

5.2 The Choice of the Mesh Spacing h

Suppose we have a good value of the penalty weighting parameter α. For a fixed sample $x_1,\ldots,x_N$ we choose the mesh nodes t_2 and t_{m-1}. Recall the estimate is zero outside the interval (t_2,t_{m-1}) by (5.1.1). We consider the resulting continuous piecewise linear discretized maximum penalized estimate as a function of the mesh interval width h. The choice of h is important since the amount of work required by the algorithm is approximately proportional to $(m-3)^3$. However, we wish to pick h sufficiently small to reveal the fine structure in the estimate.

To illustrate the practical aspects of the preceding discussion, we consider a numerical example. A random sample of size 100 was generated from the $N(0,1)$ density. The choice $\alpha = 10$ is good, as we demonstrate

in our simulation study in section 5.5. The interval of support (t_2, t_{m-1}) was taken to be (-4,4). In Diagrams 5.2.1-5.2.4 we graph the N(0,1) density (* on graph) with the discretized maximum penalized likelihood solution (0 on graph) for the choices $h = 2.0,\ 1.0,\ 0.5$ and 0.25 with corresponding values $m = 7,\ 11,\ 19$ and 35. The stability of the estimates is apparent. For this sample of size 100, choosing a smaller h does not appear to be warranted.

We remark that as h decreases, the size of our problem increases. In general, better initial guesses are required in Newton's method for larger problems than for smaller problems. If convergence problems are encountered as the result of poor initial guesses (obtained from a histogram or kernel estimate), we bootstrap the algorithm to provide good initial estimates. First a large mesh interval is chosen so that the problem is small and Newton's method converges quickly. This coarse estimate is then used to provide initial guesses for a finer mesh, say, twice as many nodes. We continue refining the mesh until h is as small as desired. Since this procedure provides excellent initial guesses, only a few iterations should be required for (5.1.9) to be satisfied.

A numerical study indicates that the D.M.P.L.E. is stable for fixed N as $h \to 0$. We have seen that this limit is precisely the monospline estimator of de Montricher, Tapia, and Thompson [1975]. It appears that among our sufficient conditions for consistency [(1) $\alpha > 0$, (2) $N \to \infty$, (3) $\lim_{N\to\infty} h^{2r}N = \infty$, (4) $\lim_{N\to\infty} h = 0$] the condition $\lim_{N\to\infty} h^{2r}N = \infty$ is an artifact of our proof. At this point it would seem that necessary and sufficient conditions for consistency of the D.M.P.L.E. are simply 1, 2, and 4. Since for fixed N the D.M.P.L.E. solutions converge to the infinite dimensional M.P.L. solution as $h \to 0$, it appears that necessary and sufficient conditions for consistency for the M.P.L.E. are one and two above.

DIAGRAM 5.2.1. N = 100 N(0,1) D.M.P.L.E. $\alpha = 10$ h = 2.0

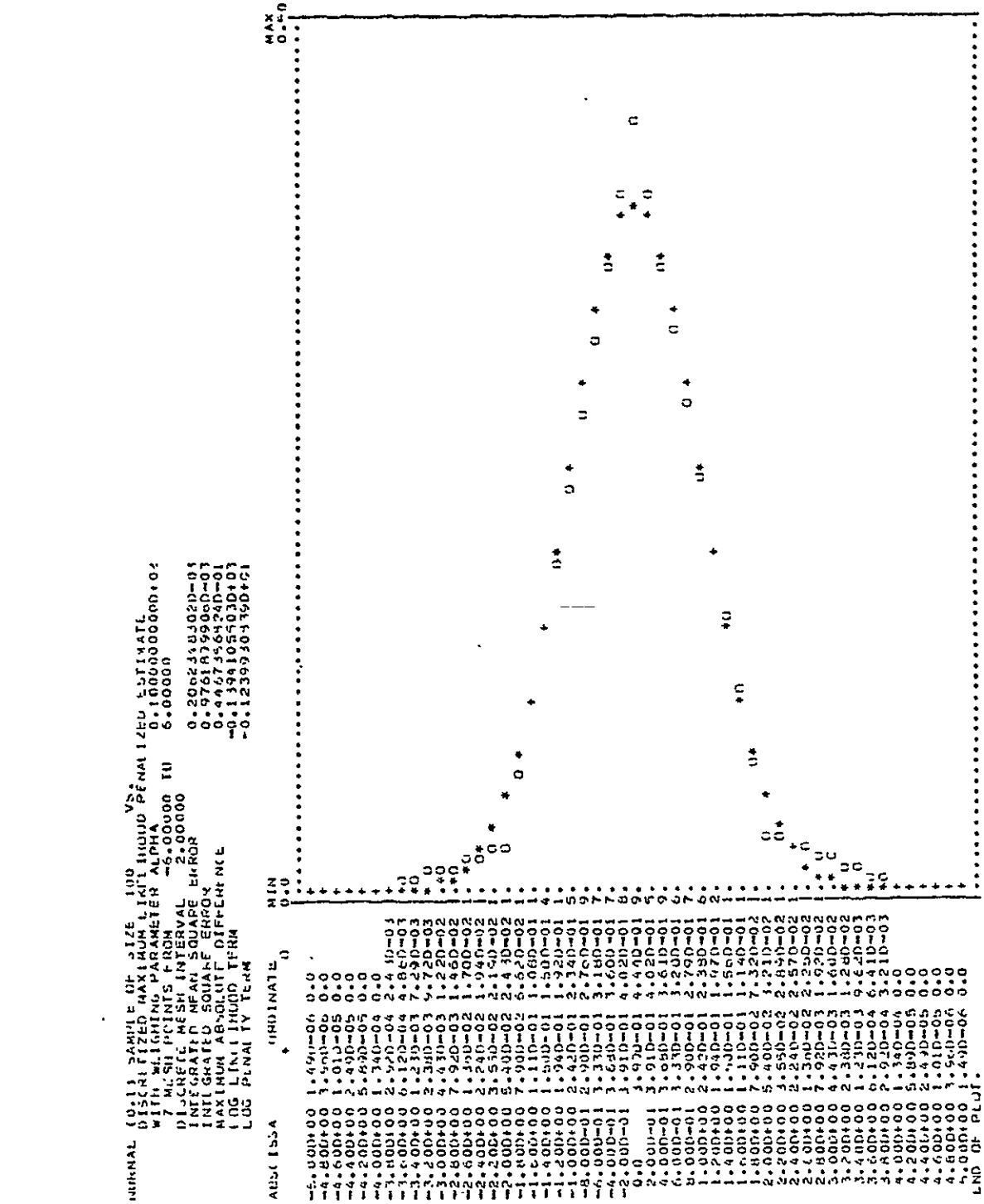


DIAGRAM 5.2.2. N = 100 N(0,1) D.M.P.L.E. α = 10 h = 1.0

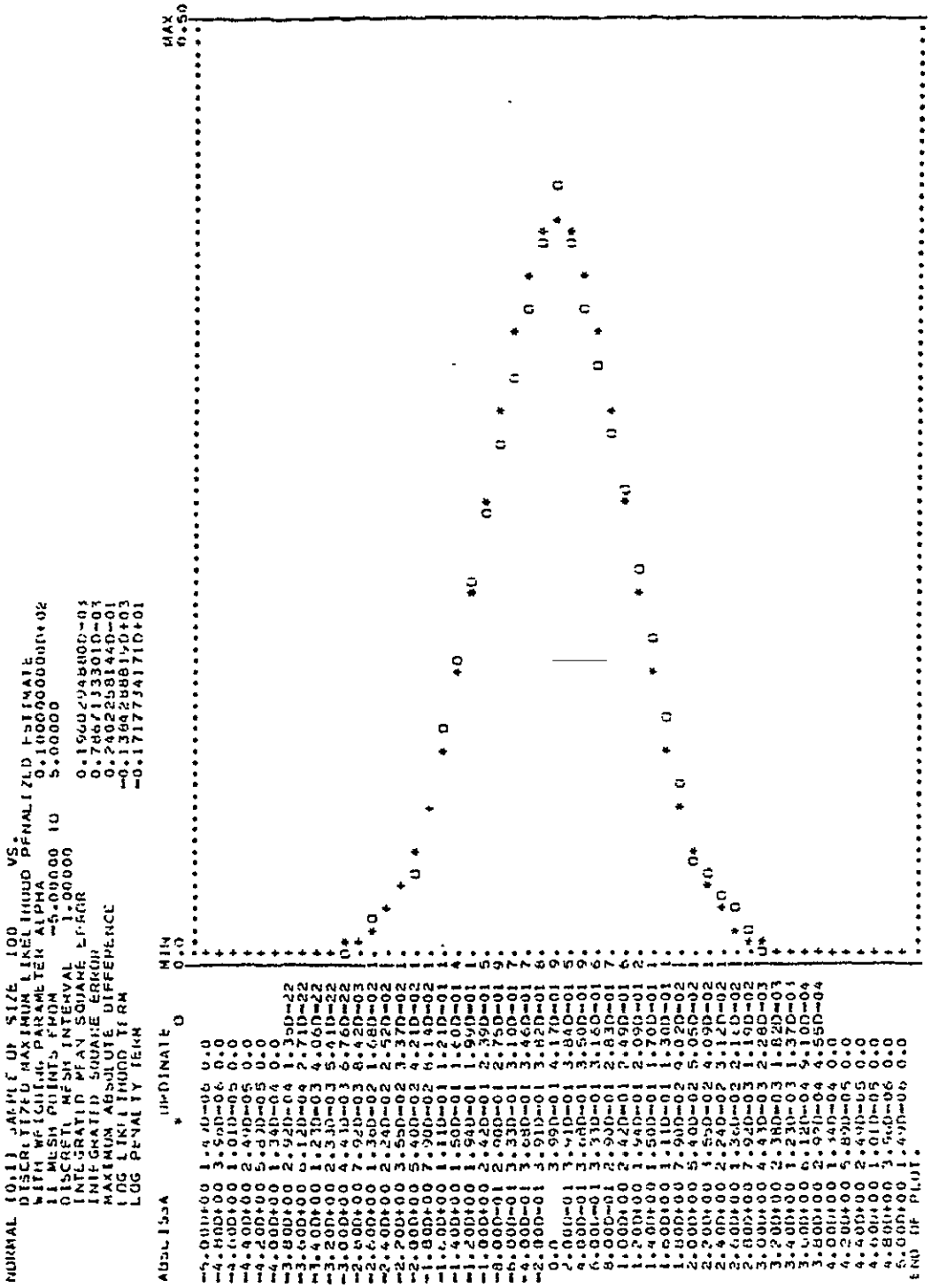


DIAGRAM 5.2.3. N = 100 N(0,1) D.M.P.L.E. α = 10 h = 0.5

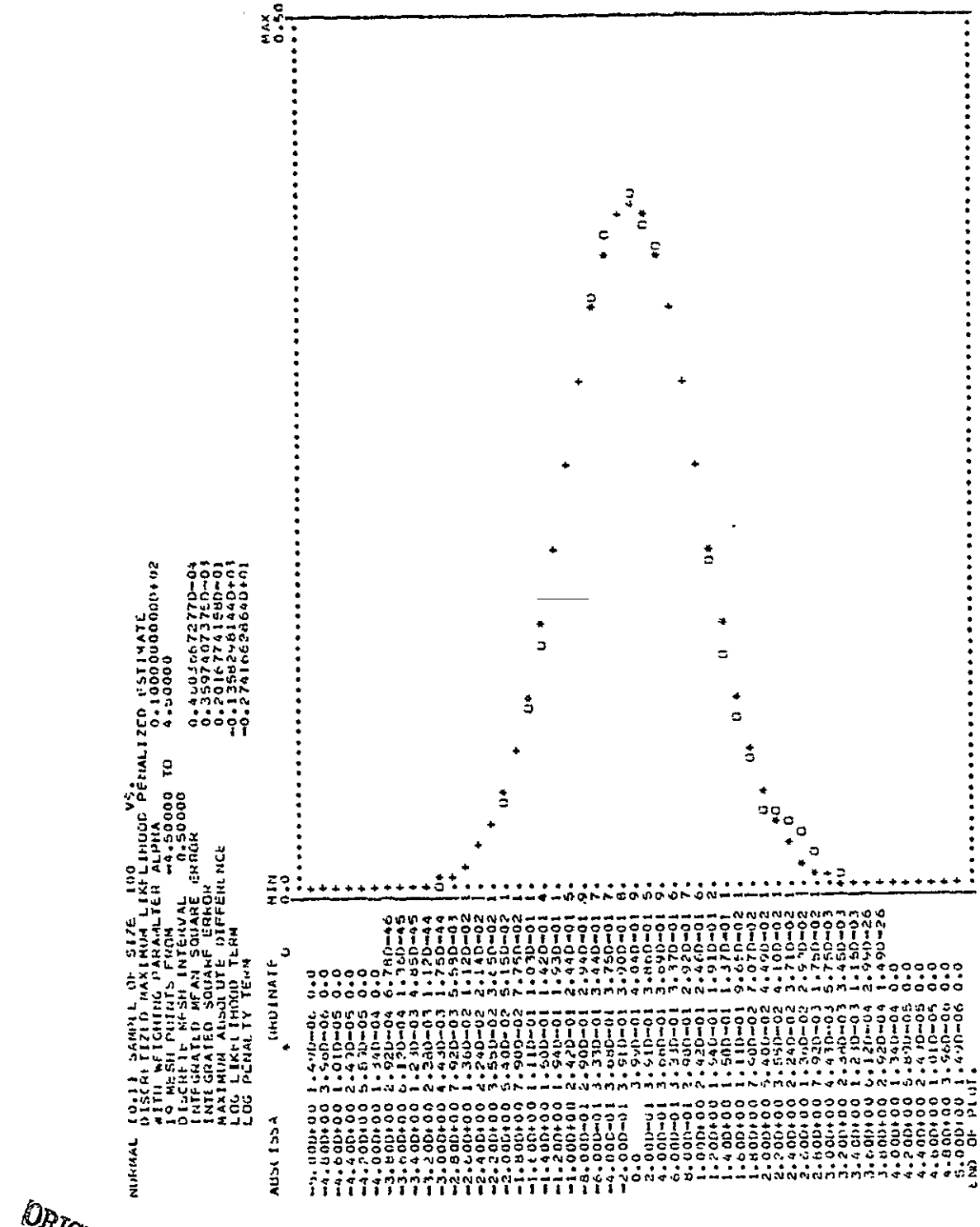
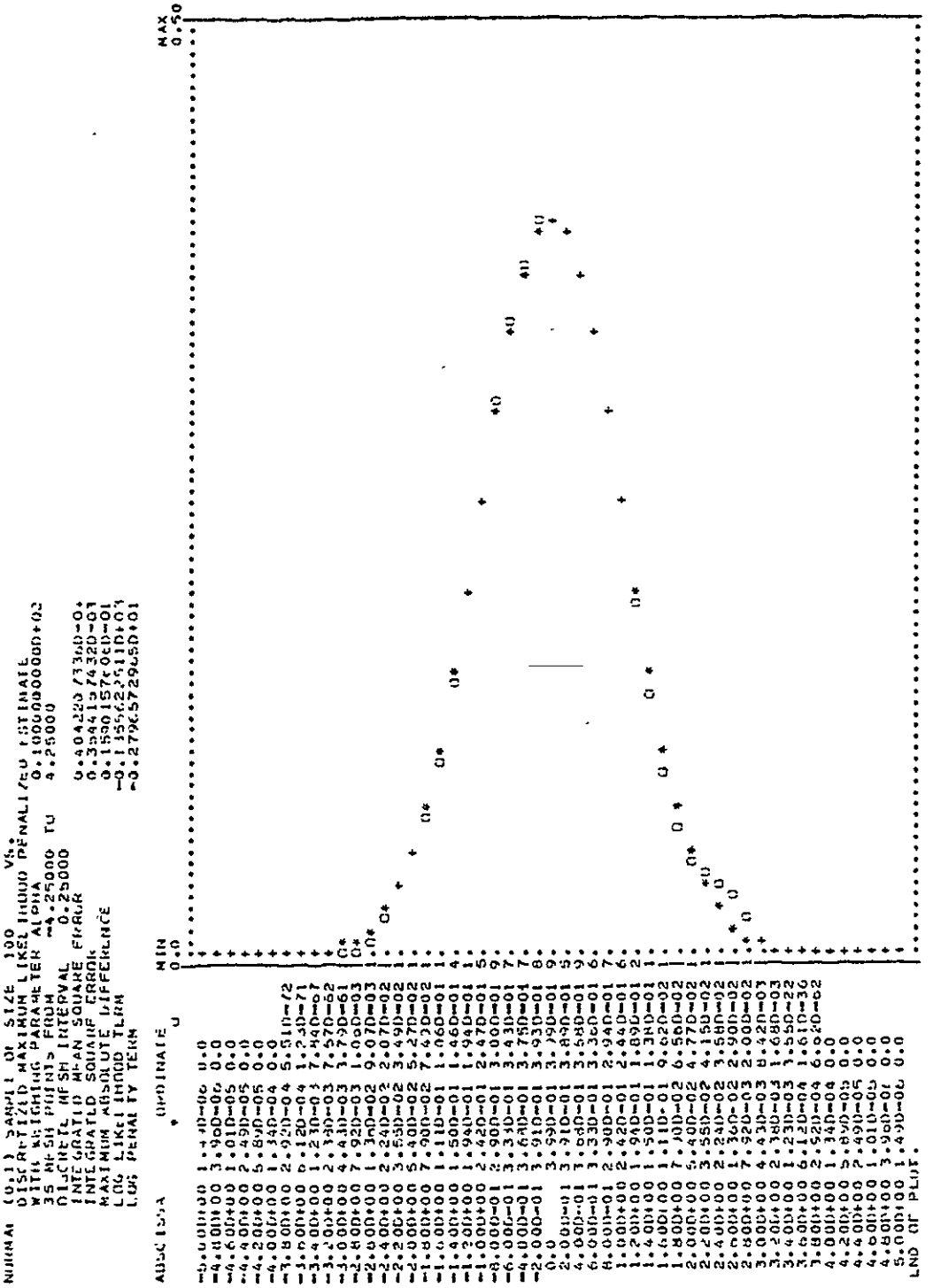


DIAGRAM 5.2.4. N = 100 N(0,1) D.M.P.L.E. α = 10 h = 0.25



5.3 The Choice of α

We next consider the choice of α. This problem is more important and more difficult than the selection of the mesh spacing h. Philosophically there is a correspondence between α and the Parzen kernel scaling parameter $h(N)$. As we discussed in Chapter 2, $h(N)$ too large results in a disperse estimate while $h(N)$ too small results in a highly varying estimate. The α parameter has the same effect for the piecewise linear estimate. To pick an appropriate α several values should be examined, picking α as small as possible without incurring a large variance in the corresponding estimate. This interactive mode is useful in practice. We hope to automate this choice of α in a manner similar to the quasi-optimal procedure for kernel estimates discussed in section 2.6.

To demonstrate graphically the discussion in the previous paragraph, a random sample of size 300 was generated from the bimodal mixture density $3/4\ N(-1.5,1) + 1/4\ N(1.5, 1/9)$. The interval (t_2, t_{m-1}) was taken as $(-5, 2.6)$ with mesh spacing $h = 0.2$ and $m = 41$. In Diagrams 5.3.1-5.3.6 the bimodal density is graphed with the solutions corresponding to $\alpha = 10^3, 10^2, 10, 1, 10^{-1}$ and 10^{-2}. Biased by the knowledge of the true underlying density we might accept the variance in the estimate with $\alpha = 0.1$, but would otherwise probably choose $\alpha = 1.0$. Even with the fixed mesh and $h = 0.2$, the variance of the estimate corresponding to $\alpha = 0.01$ is readily apparent.

The kernel estimator with the quartic kernel given in Table 2.5.1 was applied to the same bimodal data for a sequence of values of the scaling factor $h(N)$. For α and $h(N)$ too small the corresponding estimates have sharp peaks. As $h(N) \to 0$, the kernel estimate approximates the delta function solution; however, as $\alpha \to 0$, the discrete solution cannot

come arbitrarily close to the delta functions because the mesh interval is a fixed, positive number. As $h(N) \to \infty$, the kernel estimate approximates a diffuse uniform density that retains little semblance of the true sampling density; however, as $\alpha \to \infty$, the discrete solution looks like Diagram 5.3.1 since the mesh interval and the support interval (t_2, t_{m-1}) are fixed. Notice that the bimodal nature of the samples is apparent even for the over-smoothed estimate ($\alpha = 10^3$). Thus the choice of a mesh makes the discrete solution more robust than the kernel estimate with respect to the parameters α and $h(N)$.

The Interactive Mode

We recommend the following procedure for applying the penalized likelihood algorithm to a given sample $x_1, \ldots, x_N$. The range of the sample is examined and any outliers truncated if desired. A histogram is useful in this aspect. A good estimate of the penalty weighing factor α can be obtained with a coarse mesh as demonstrated in Diagrams 5.2.1-5.2.4. Therefore, we choose a large value of the mesh interval h and try various values of α in powers of ten. Then we pick α as small as possible in accordance with our prior feelings about the variance of the resulting estimate. For initial guesses we use the histogram estimate or one-hundredth, whichever is greater. For the Lagrange multiplier we use $-N/4$. Once an acceptable α is found the mesh interval h is decreased until the fine structure is apparent. At this point α may be changed to further smooth or unsmooth the estimate.

DIAGRAM 5.3.1. N = 300 Bimodal D.M.P.L.E. $\alpha = 10^3$

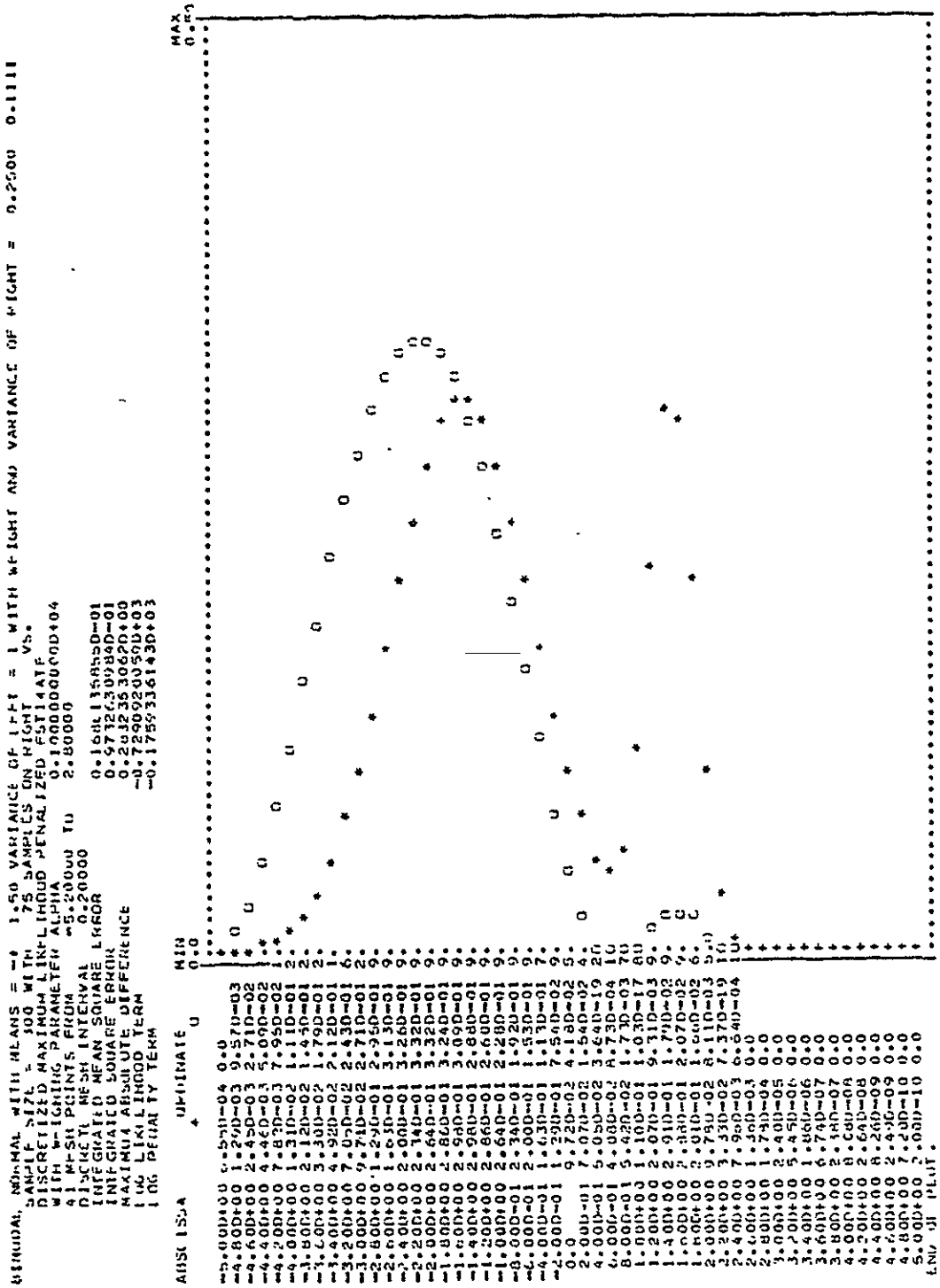


DIAGRAM 5.3.2 N = 300 Bimodal D.M.P.L.E. $\alpha = 10^2$

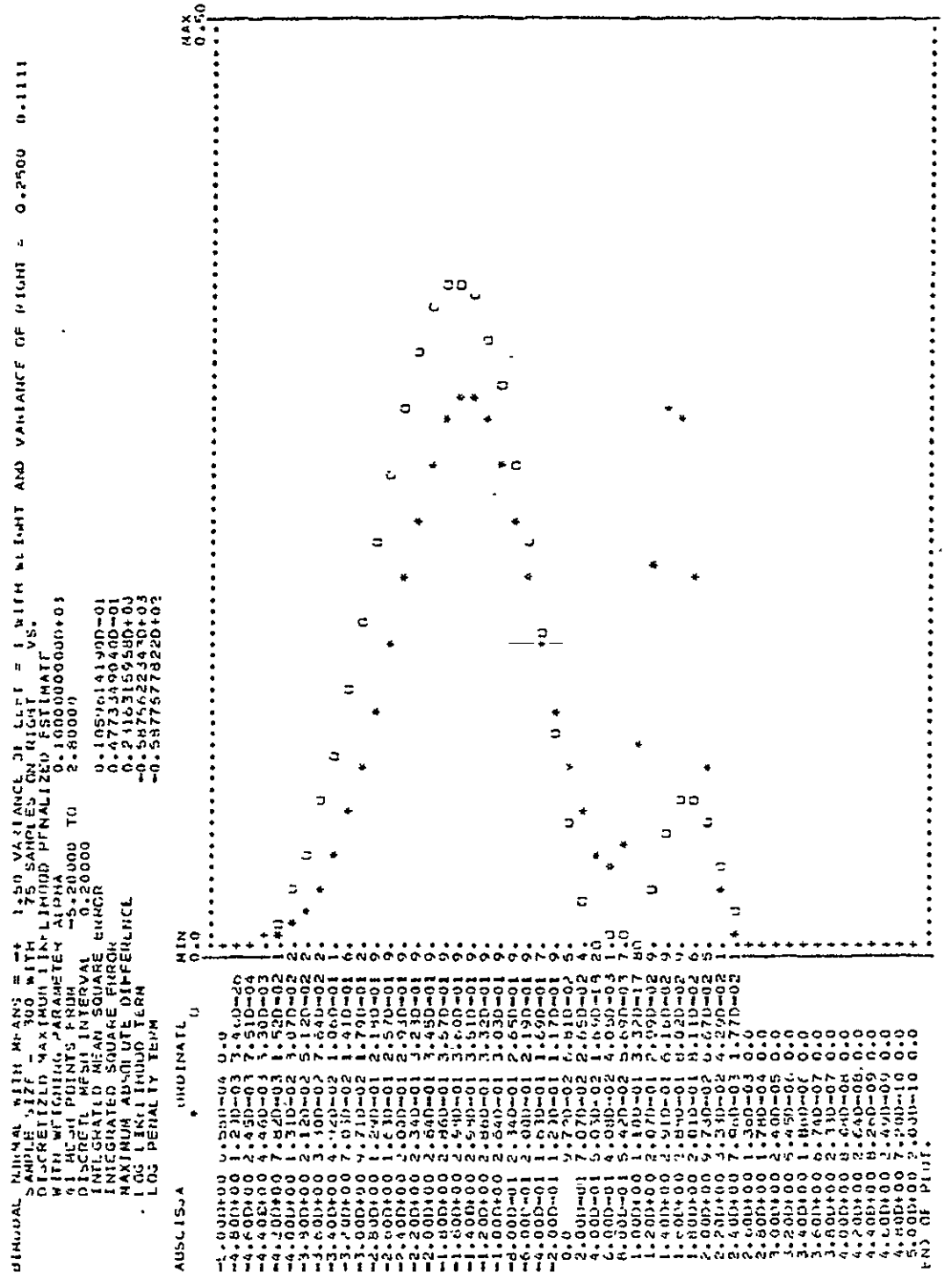


DIAGRAM 5.3.3 N = 300 Bimodal D.M.P.L.E. $\alpha = 10$

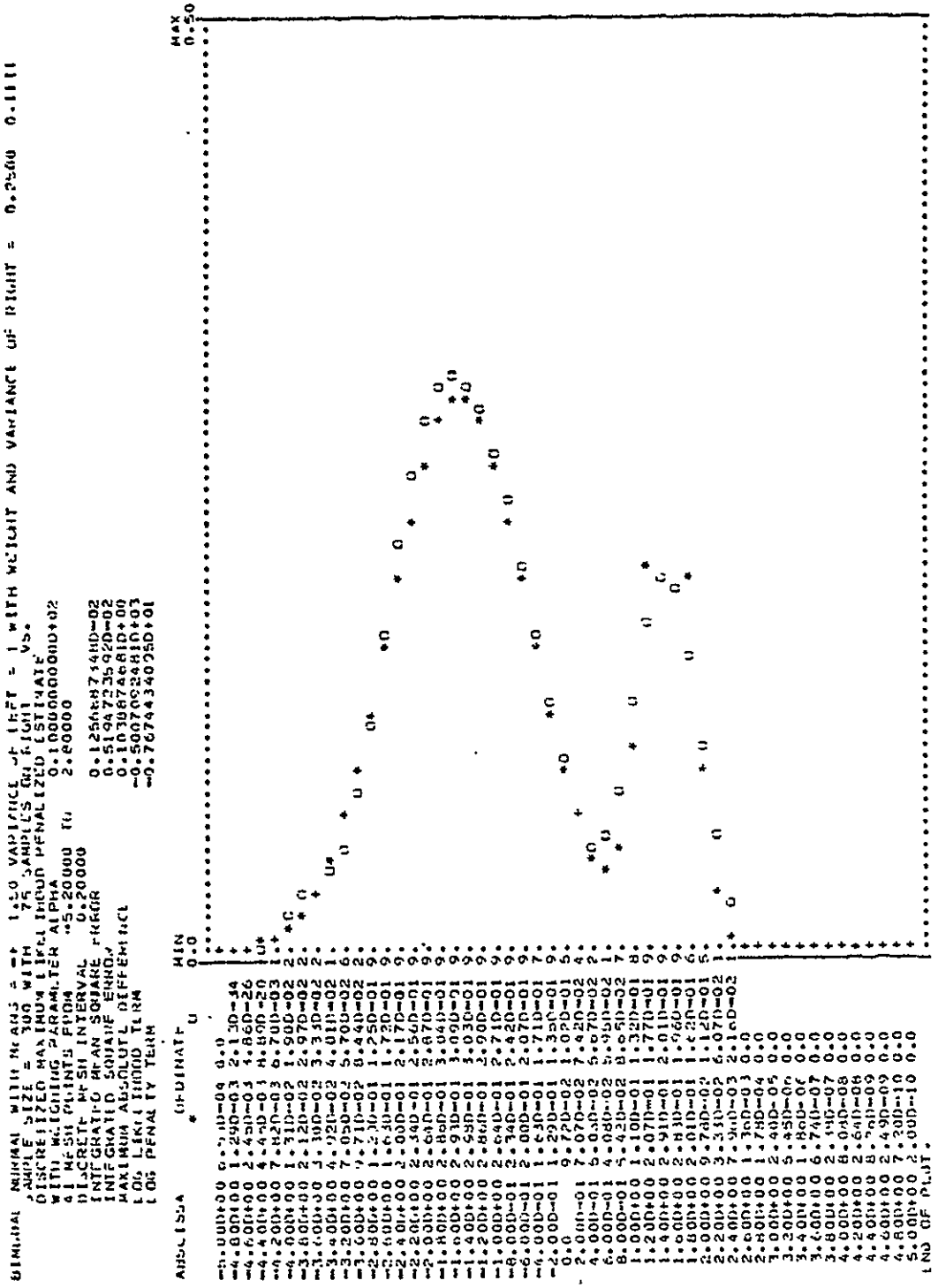
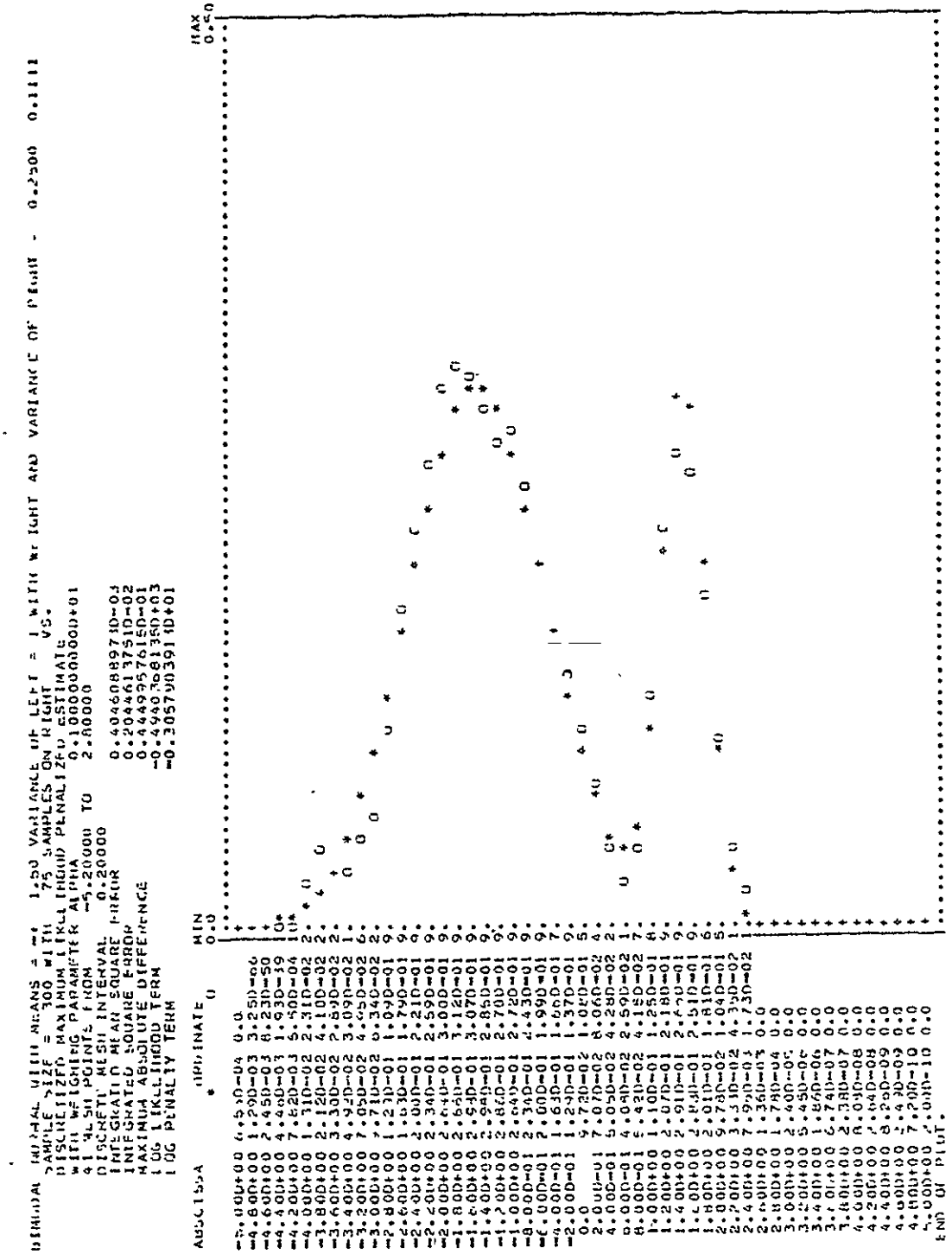


DIAGRAM 5.3.4. N = 300 Bimodal D.M.P.L.E. $\alpha = 1$



C.2

DIAGRAM 5.3.5 N = 300 Bimodal D.M.P.L.E. $\alpha = 10^{-1}$

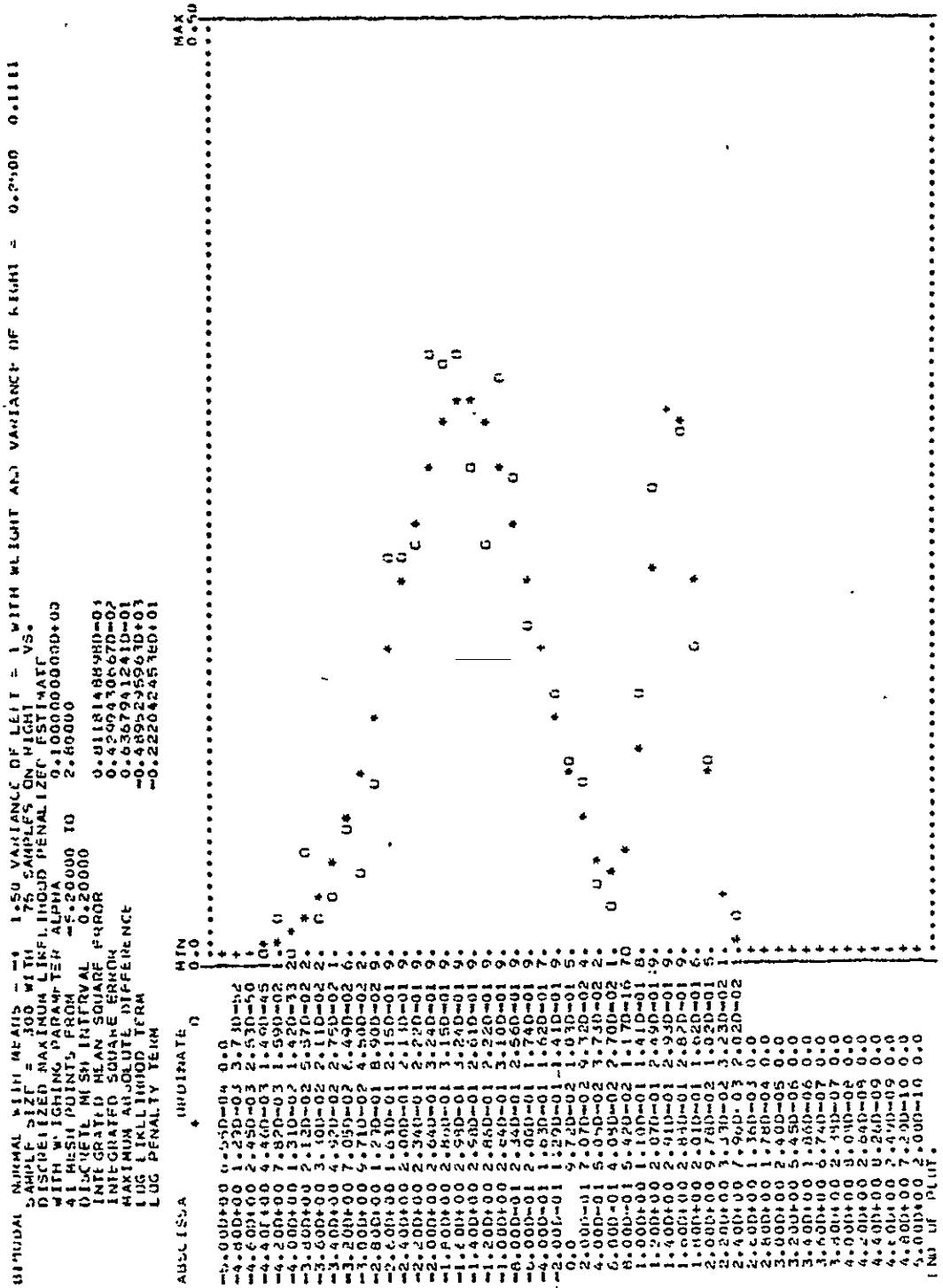


DIAGRAM 5.3.6. N = 300 Bimodal D.M.P.L.E. $\alpha = 10^{-2}$

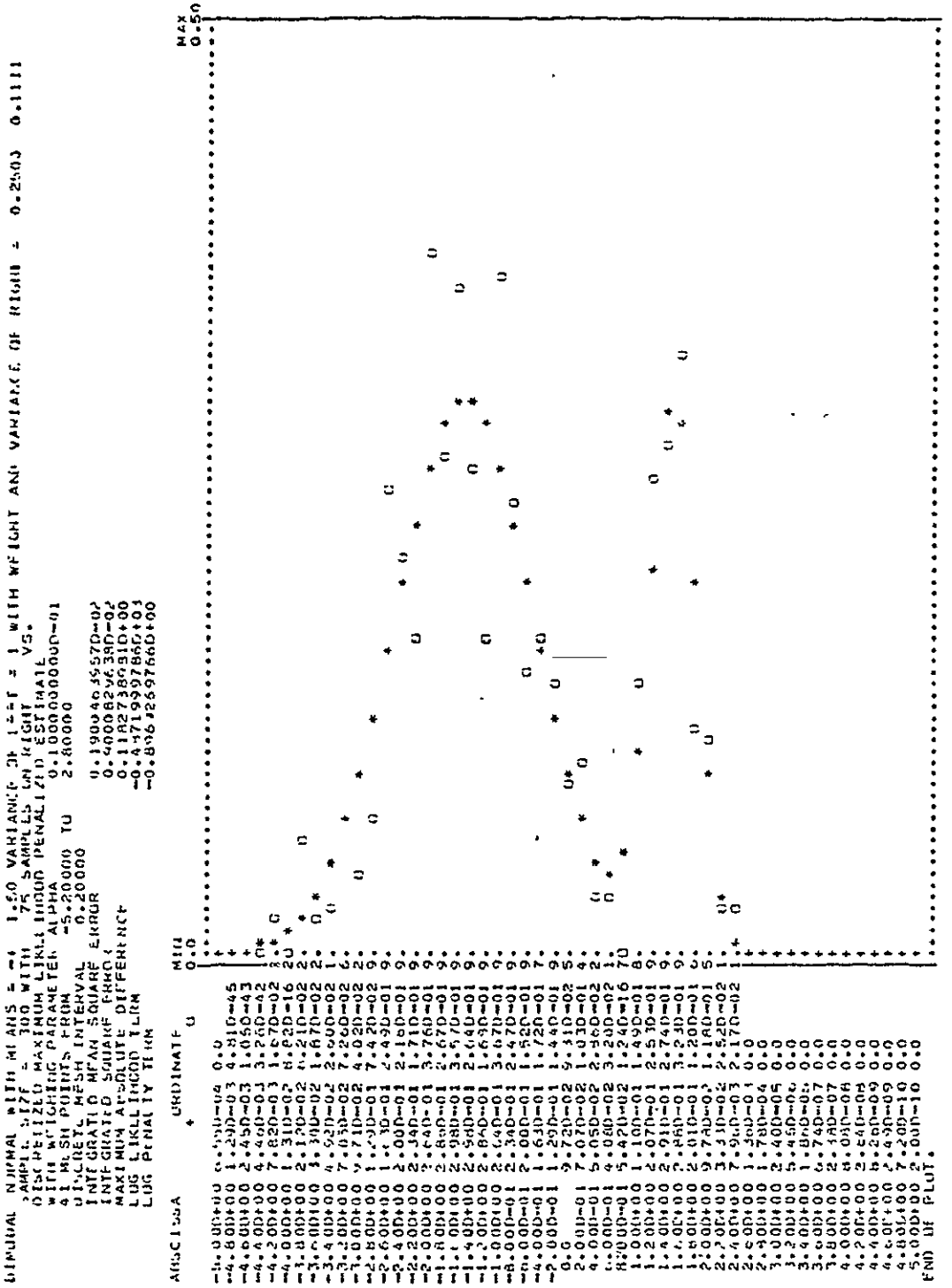


DIAGRAM 5.3.7. N = 300 Bimodal Quartic Kernel h(N) = 5.0

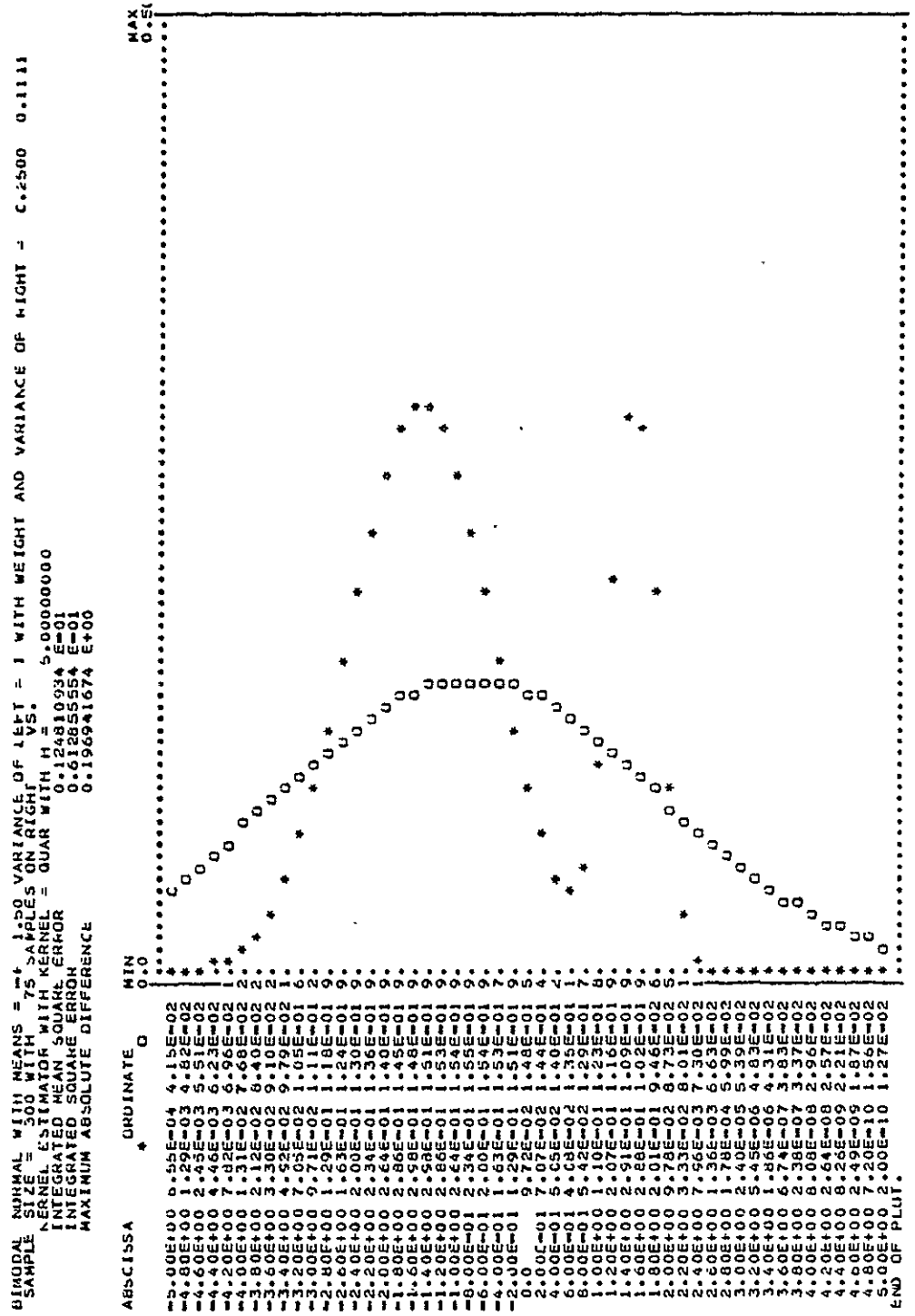


DIAGRAM 5.3.8. N = 300 Bimodal Quartic Kernel h(N) = 2.0

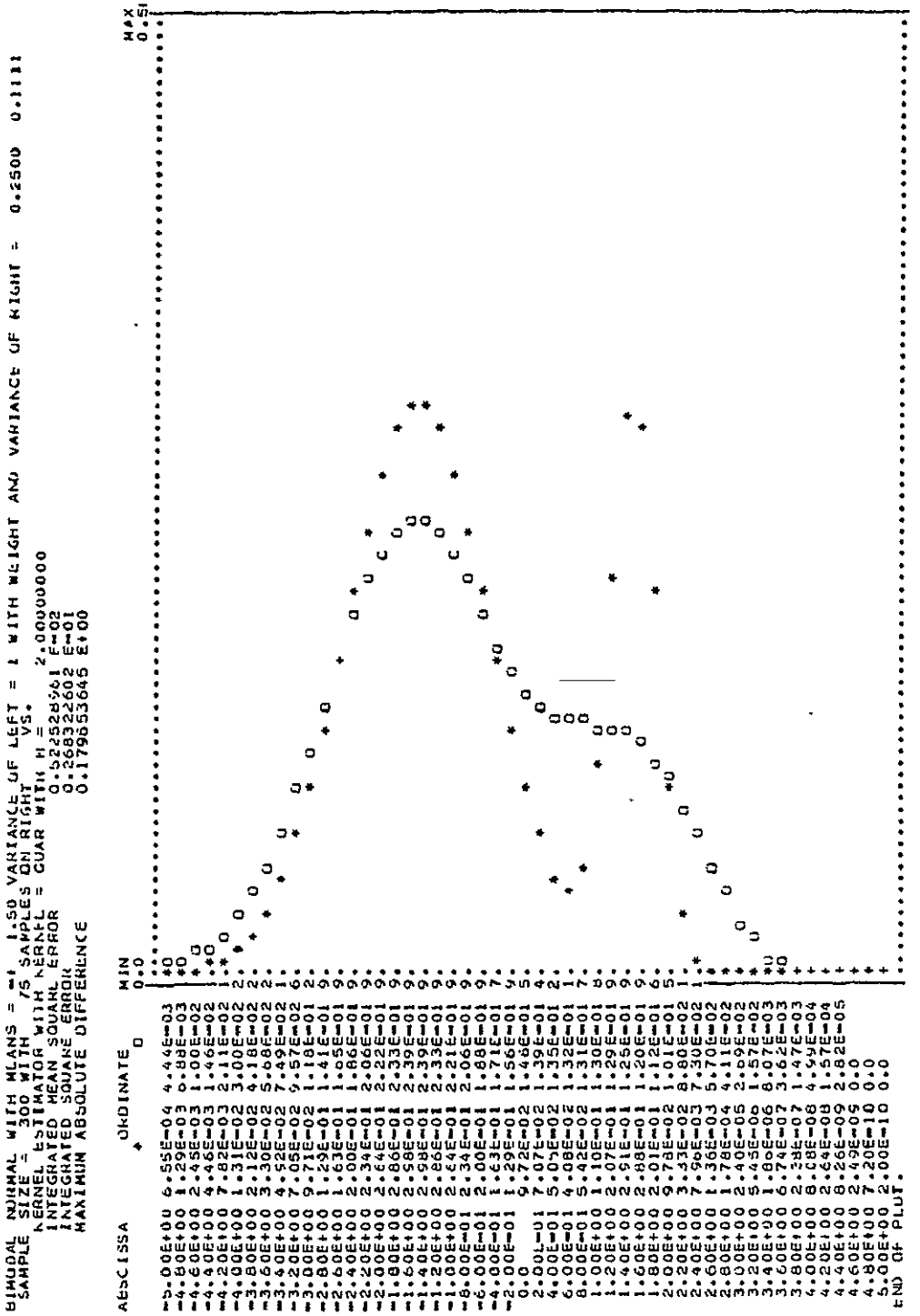


DIAGRAM 5.3.9. N = 300 Bimodal Quartic Kernel h(N) = 0.8

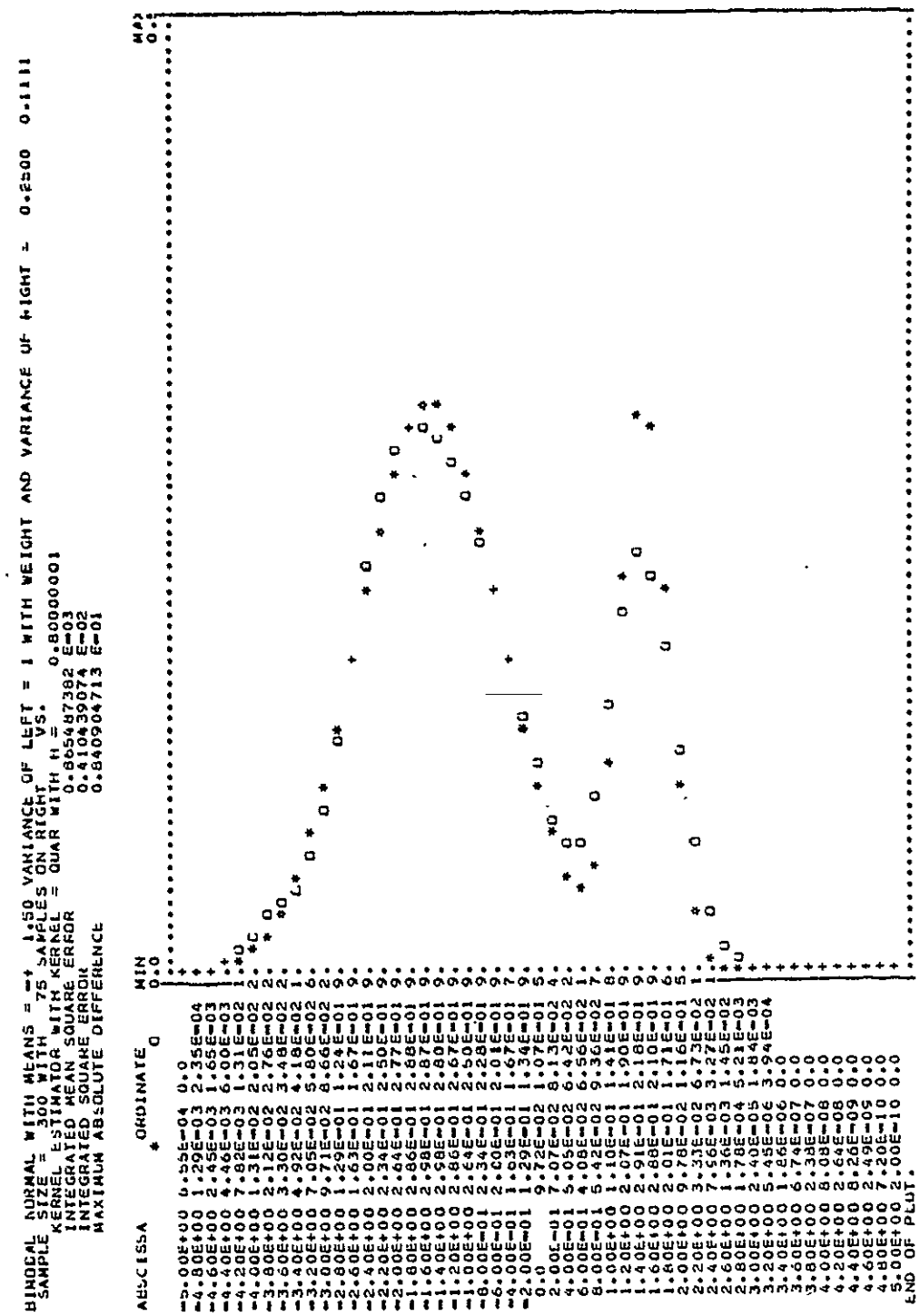


DIAGRAM 5.3.10. N = 300 Bimodal Quartic Kernel h(N) = 0.6

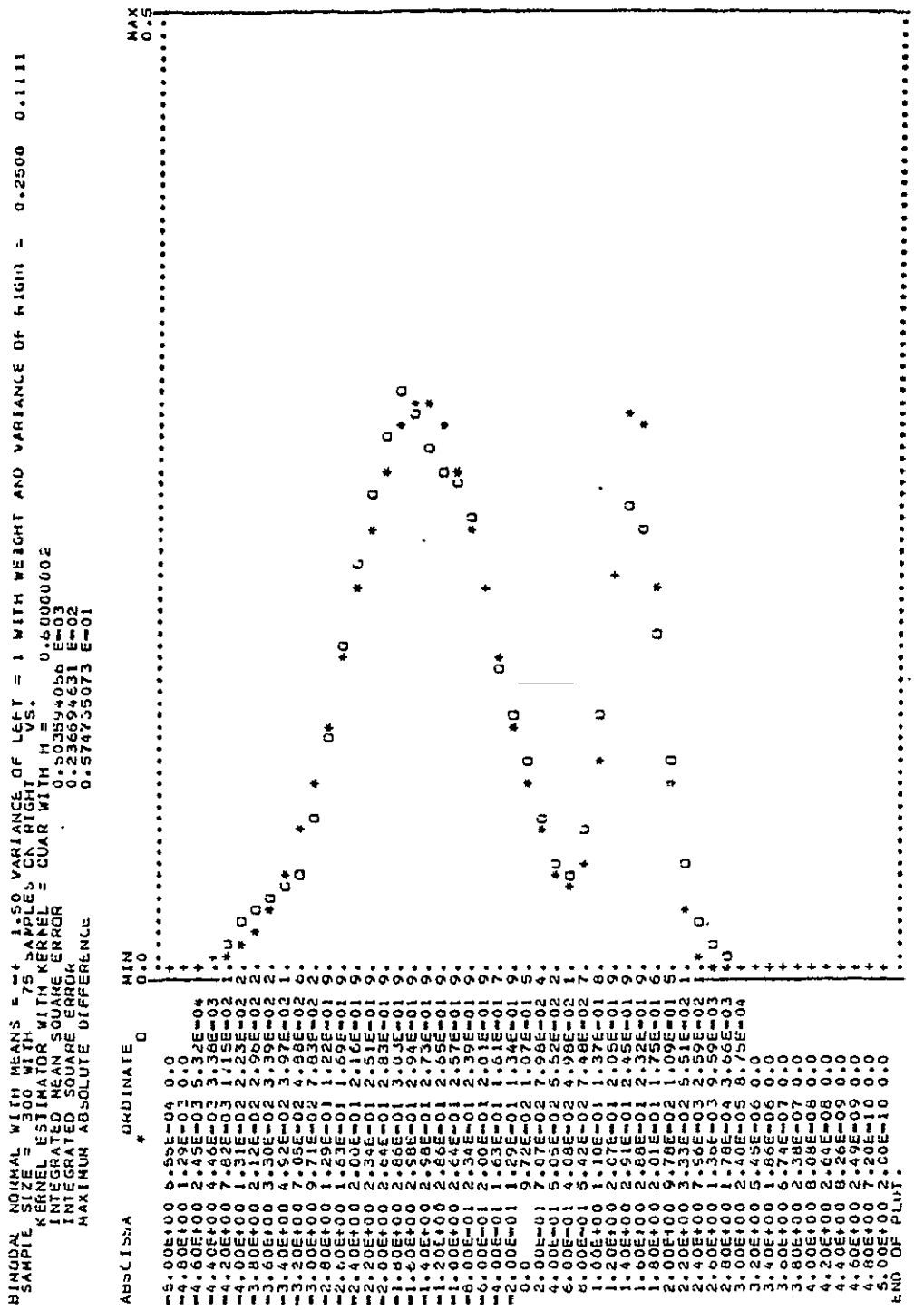


DIAGRAM 5.3.11. N = 300 Bimodal Quartic Kernel h(N) = 0.4

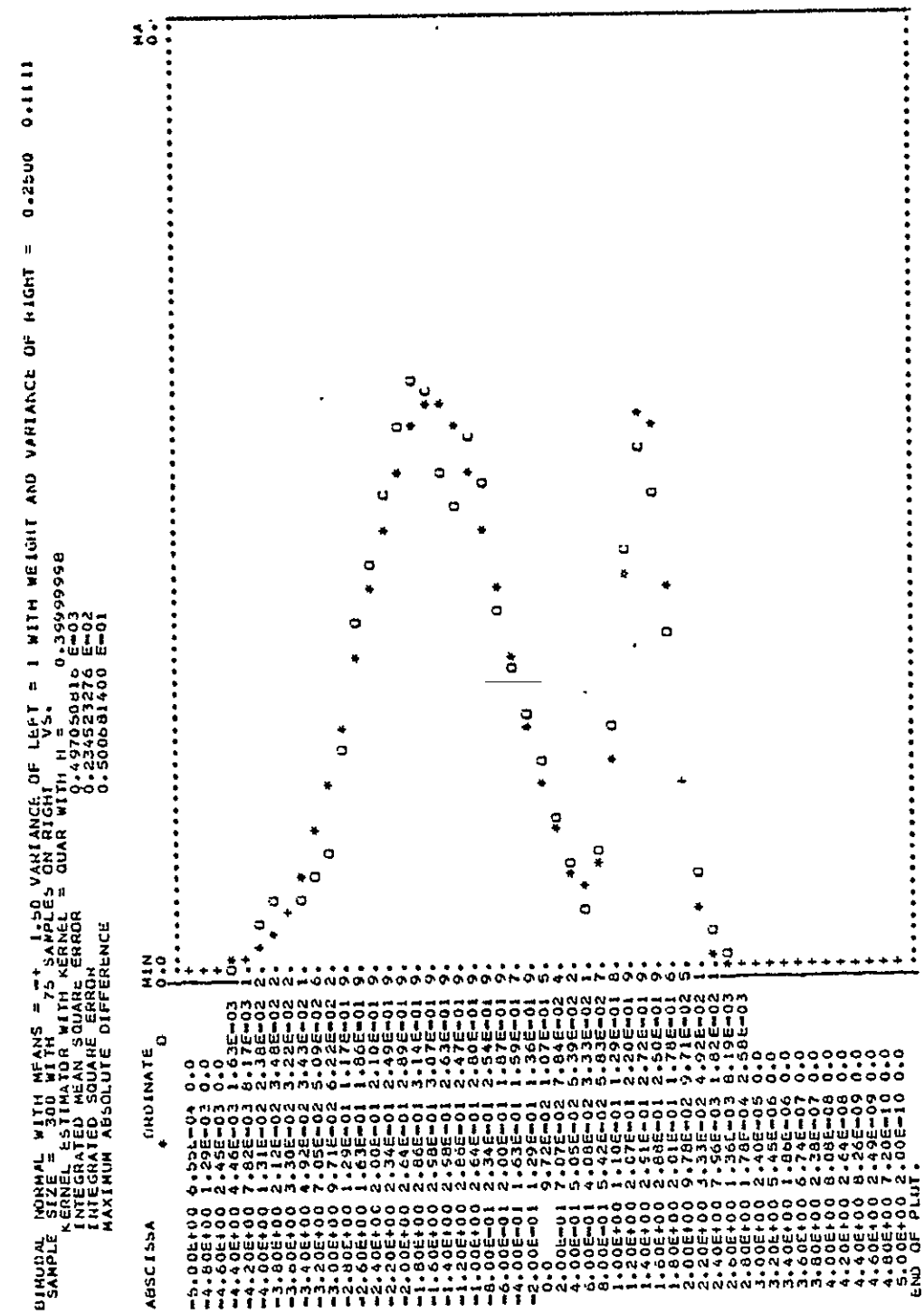
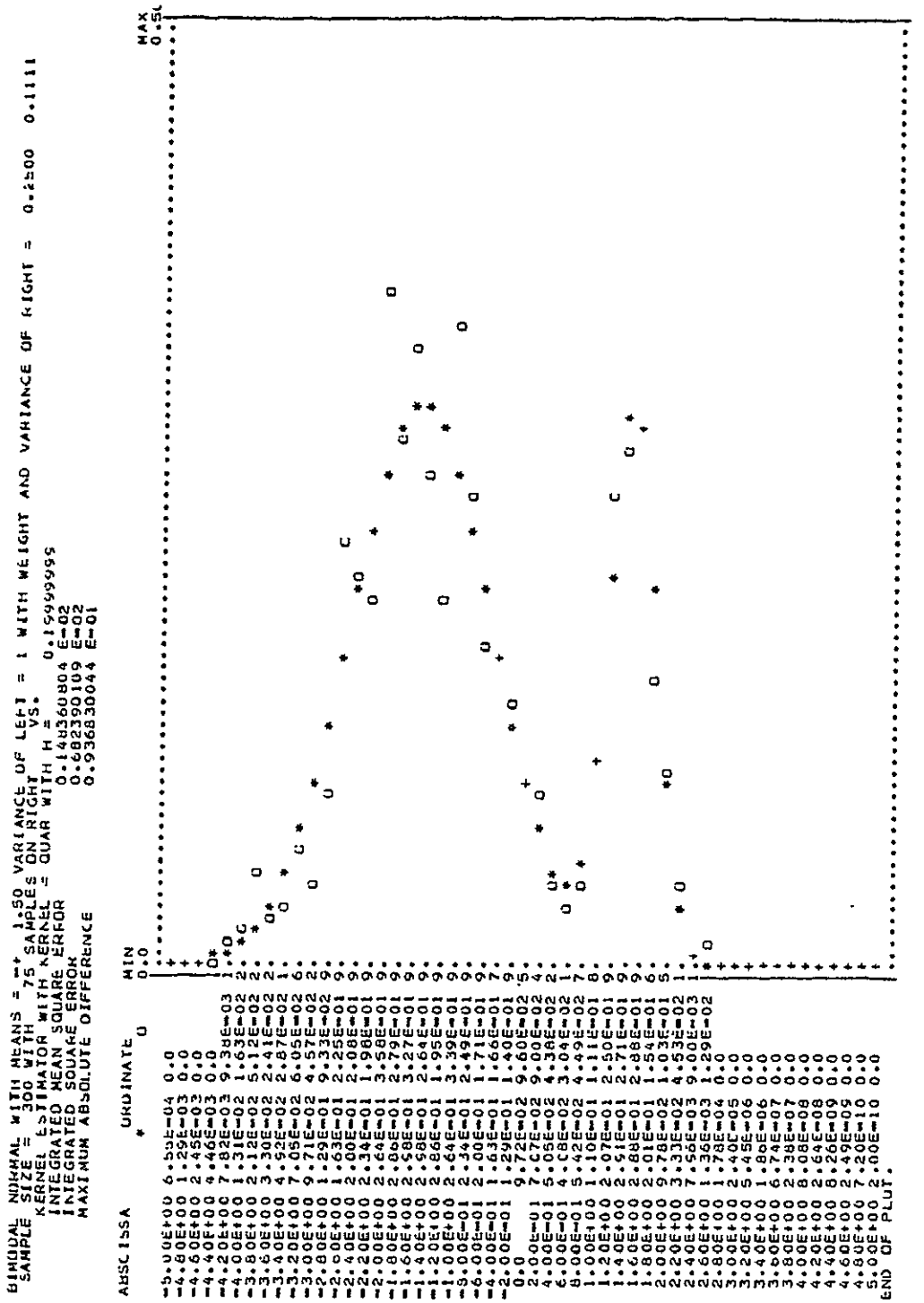


DIAGRAM 5.3.12. N = 300 Bimodal Quartic Kernel h(N) = 0.2



5.4 Examples of Kernel and Discretized Estimates

To evaluate the various estimators, four densities were chosen as benchmarks. They are:

1. the standard normal $N(0,1)$ (5.4.1)
2. bimodal $\frac{1}{2}N(-1.5,1) + \frac{1}{2}N(1.5,1)$
3. student's distribution t_5
4. the $F_{10,10}$ density shifted by 3 units for convenience.

The $N(0,1)$ density was chosen for its universal importance in sampling. The bimodal density was chosen because it sometimes occurs in situations where the standard normal is assumed; for example, the density of IQ's for U.S. high school seniors is bimodal in nature. The t_5 density was chosen for its heavy tails. Finally the $F_{10,10}$ density was chosen because it is not symmetric and has a sharp peak.

Monte Carlo simulations were performed on each of the densities in (5.4.1) using the kernel estimator and the continuous piecewise linear estimator. In Diagrams 5.4.1-5.4.7 we compare three estimates on each of several random data sets. For each random sample the discretized solution is given first (a). Then the recent non-L^1 Fourier kernel (2.1.6) of Davis [1975] is given (b). Finally, in (c) the quartic kernel (see Table 2.5.1) gives estimates indistinguishable from the Gaussian kernel. The optimal choices for the kernel scaling parameter were calculated using (2.1.3) and

$$h(N) = [1.5\sqrt{\log_{10} N}]^{-1} \qquad (5.4.2)$$

for the Fourier kernel (see section 2.1). Formula (5.4.2) was used in all cases to illustrate the practical difficulties in choosing $h(N)$ for an unknown density function. This formula is optimal for the standard normal. Even in this situation, the Fourier kernel introduces oscillations

and negative lobes in the tails of the estimate. For a random sample of size 400 from the $F_{10,10}$ density, several estimates using the Davis kernel for various choices of $h(N)$ are given in Diagrams 5.4.7b-5.4.7b'''. Yet with $N = 400$ the negative oscillations are always apparent particularly on the left where there is no probability mass (which leads to a good integrated mean square error). We remark that as a function of the scaling parameter $h(N)$ the Davis estimator behaves differently than the usual Parzen estimator. In particular, for $h(N)$ too large the resulting estimates are oversmoothed; however, unlike other kernels, the Davis estimate has large low frequency oscillations in the tails.

The optimal value of $h(N)$ for the quartic kernel estimator was obtained using formula (2.5.1); that is, for a general kernel $K(\cdot)$ satisfying (2.1.1)

$$h(N)^5 = \frac{\int_{-\infty}^{\infty} K^2(x)dx}{\left[\int_{-\infty}^{\infty} x^2K(x)dx\right]^2 \int_{-\infty}^{\infty} f''(x)^2dx} N^{-1} \qquad (5.4.3)$$

In Table 5.4.1 we give the quantities in (5.4.3) relating to the choice of a kernel from Table 2.5.1.

TABLE 5.4.1

Kernel	$\int_{-\infty}^{\infty} K^2(x)dx$	$\left[\int_{-\infty}^{\infty} x^2K(x)dx\right]^2$
Box	1/2	1/9
Triangle	2/3	1/36
Quartic	5/7	1/49
Gaussian	$1/(2\sqrt{\pi})$	1

In Table 5.4.2 we give the quantity in (5.4.3) relating to the choice of a sampling density from (5.4.1).

TABLE 5.4.2

Sampling Density	$\int_{-\infty}^{\infty} f''(x)^2 dx$
N(0,1)	$3/(8\sqrt{\pi})$
Bimodal	$(3/(16\sqrt{\pi}))(1 - 1.25e^{-2.25})$
t_5	$143/(20\pi\sqrt{5})$
$F_{10,10}$	272160/7429

DIAGRAM 5.4.1a. N = 10 N(0,1) D.M.P.L.E. α = 10

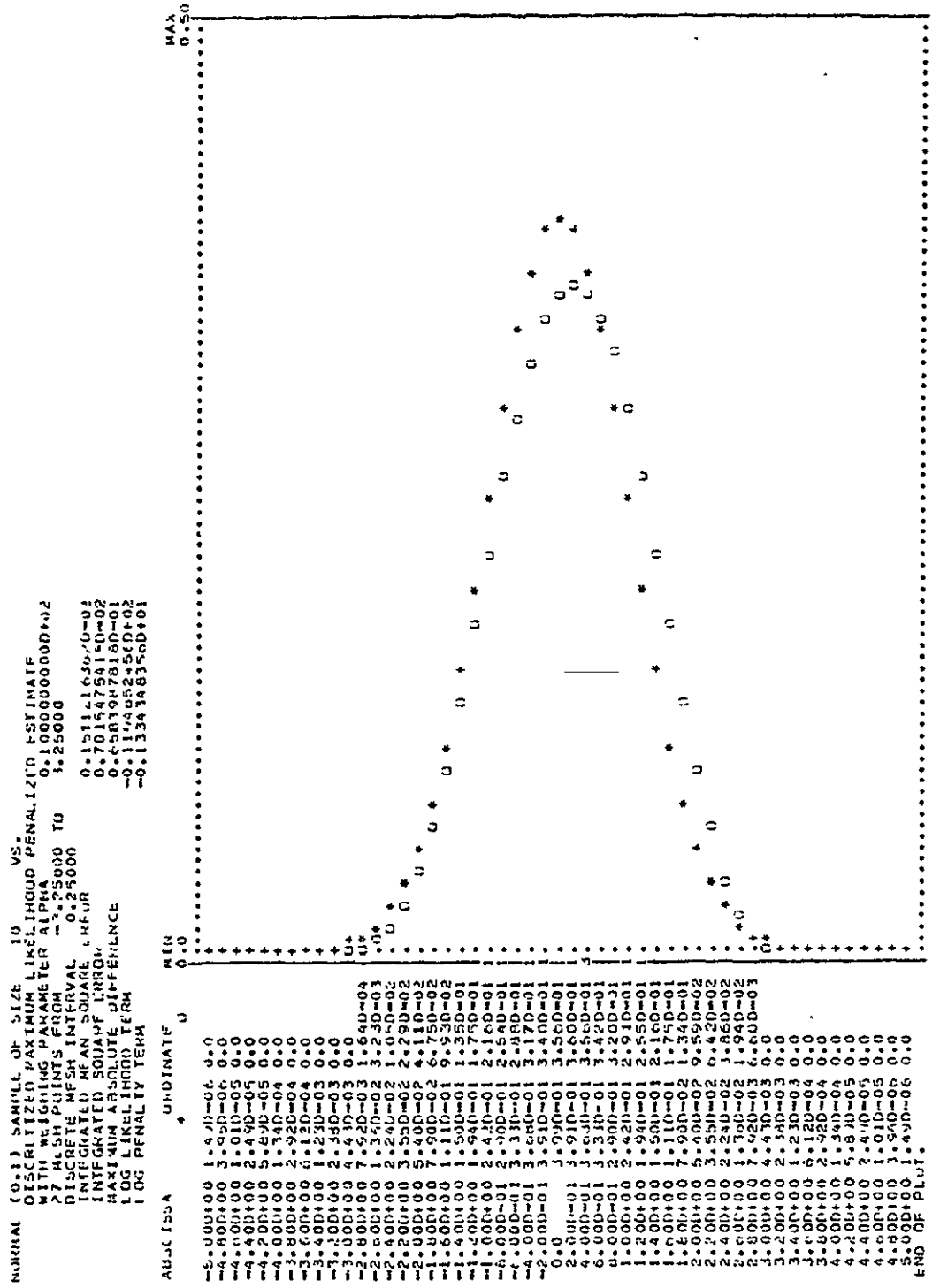


DIAGRAM 5.4.1b. N = 10 N(0,1) F.I.E. Kernel h(N) = 0.67

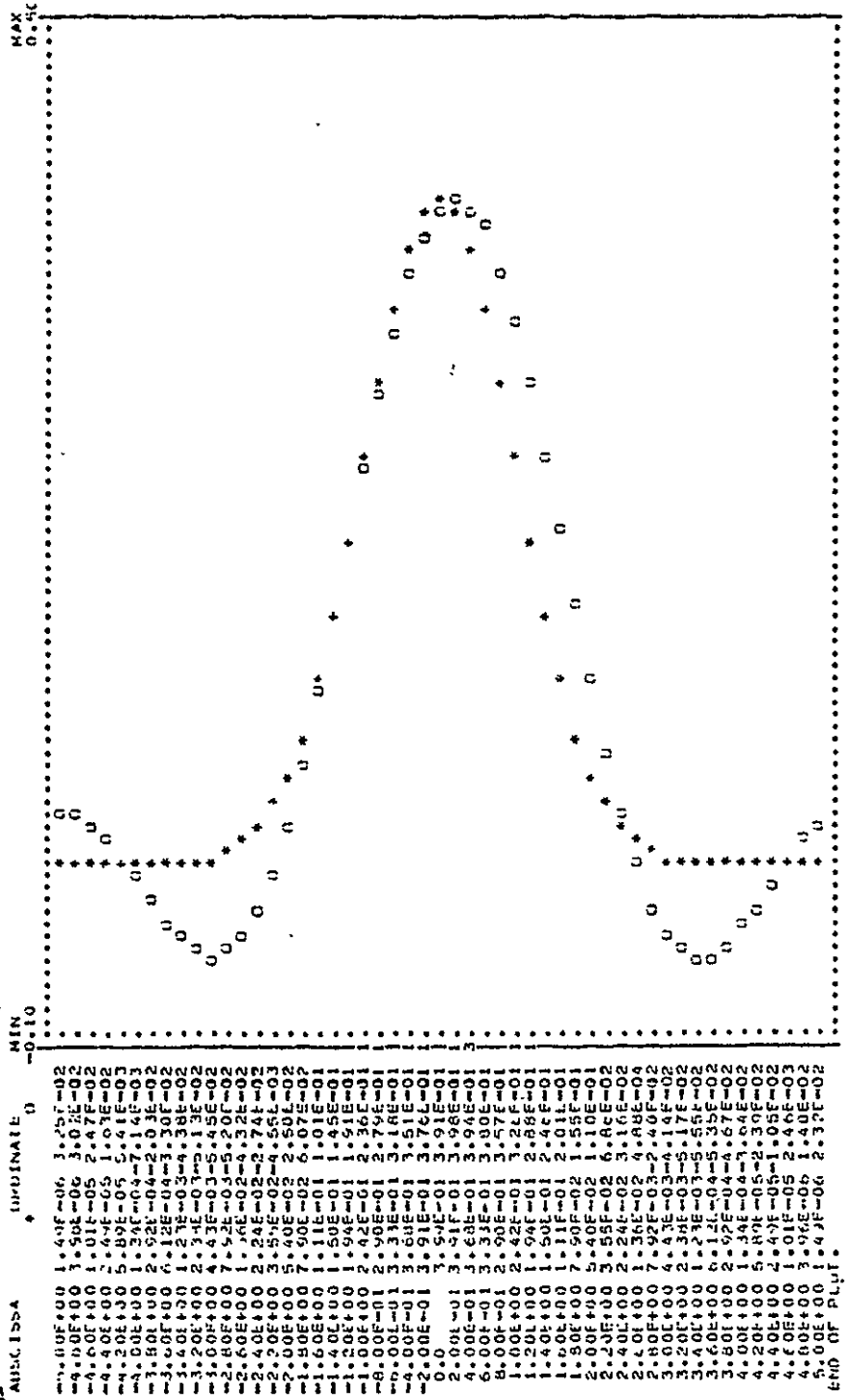


DIAGRAM 5.4.1c. N = 10 N(0,1) Quartic Kernel h(N) = 1.75

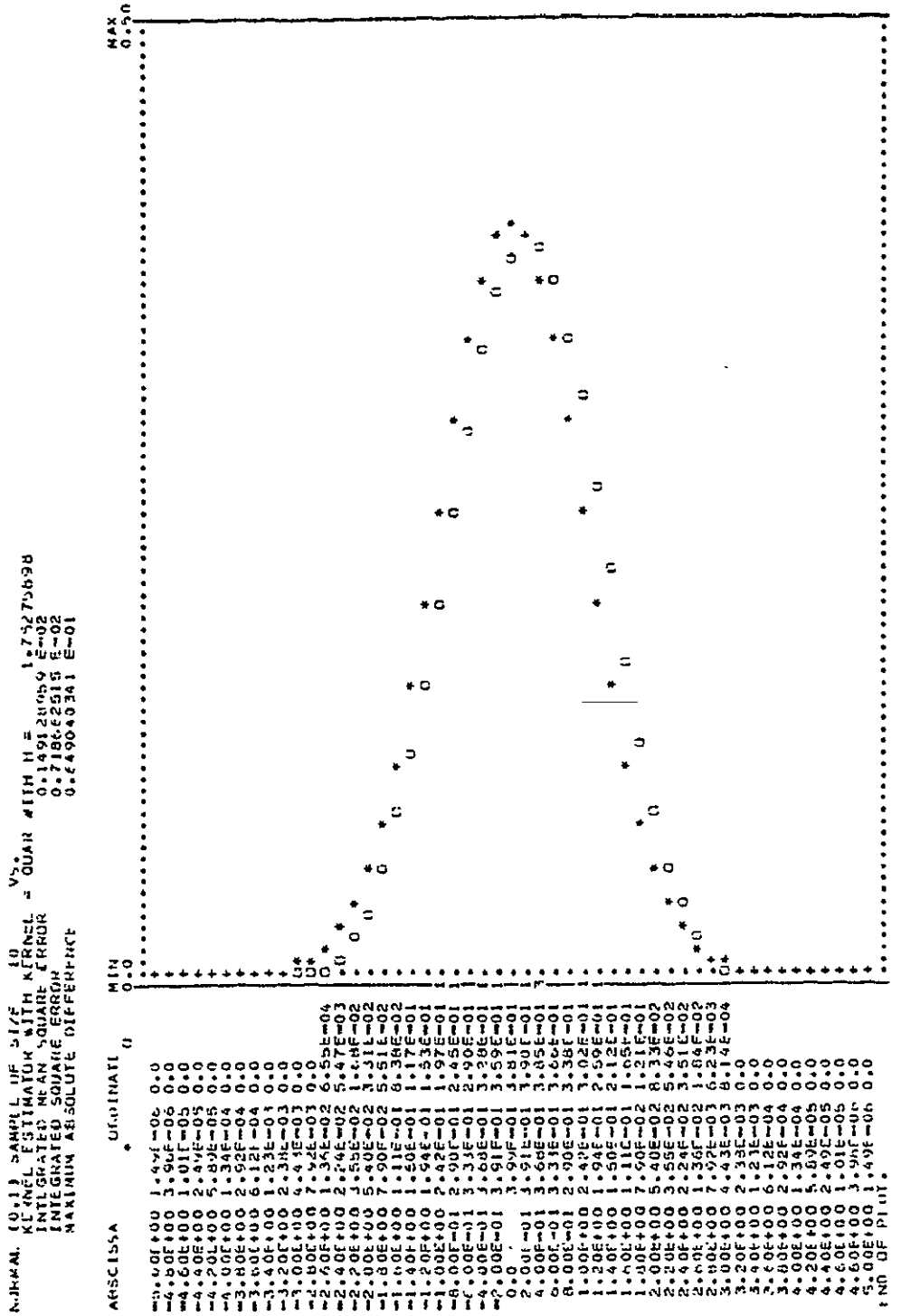


DIAGRAM 5.4.2a. N = 20 N(0,1) D.M.P.L.E. α = 10

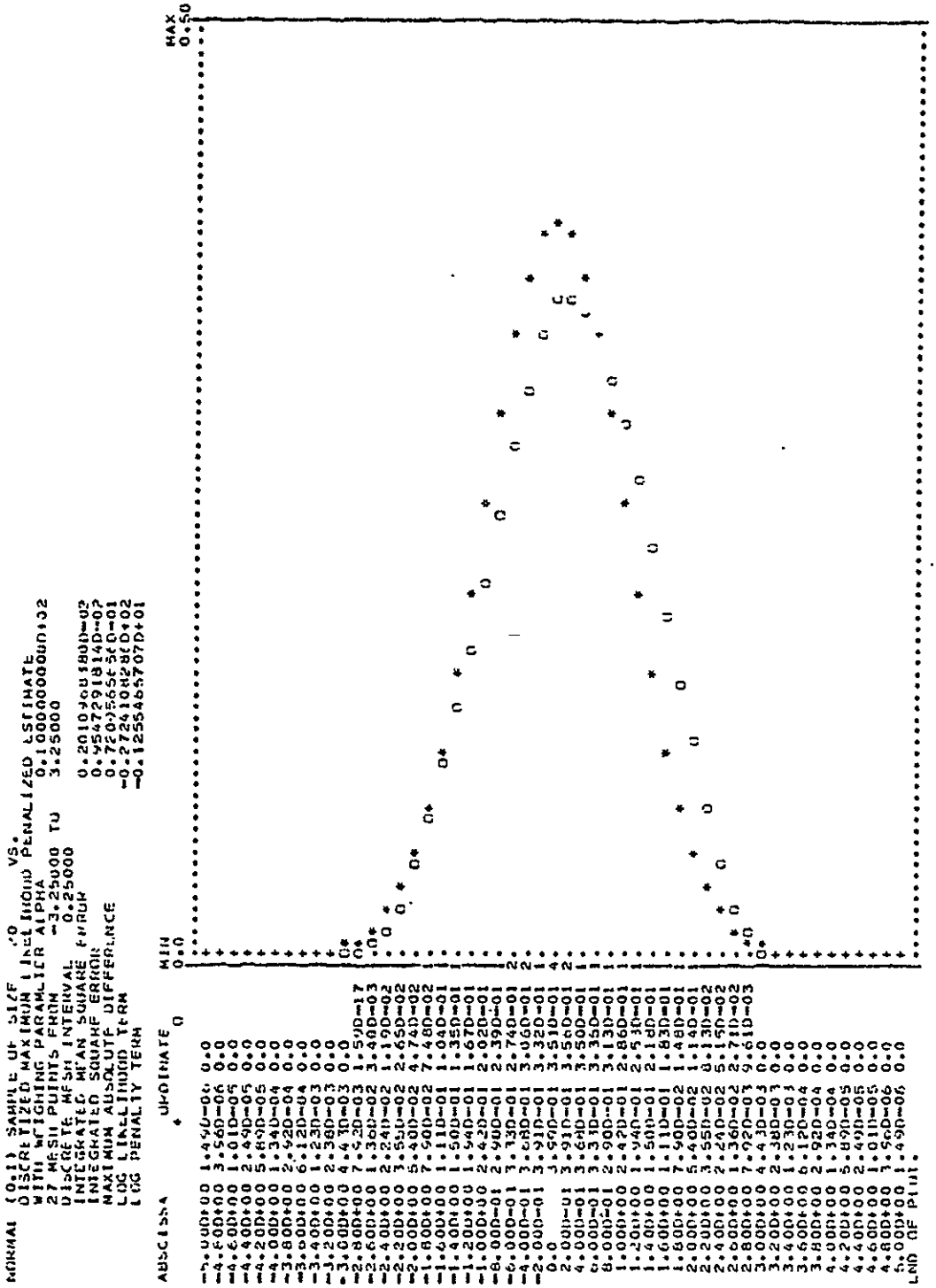


DIAGRAM 5.4.2b. N = 20 N(0,1) F.I.E. Kernel h(N) = 0.58

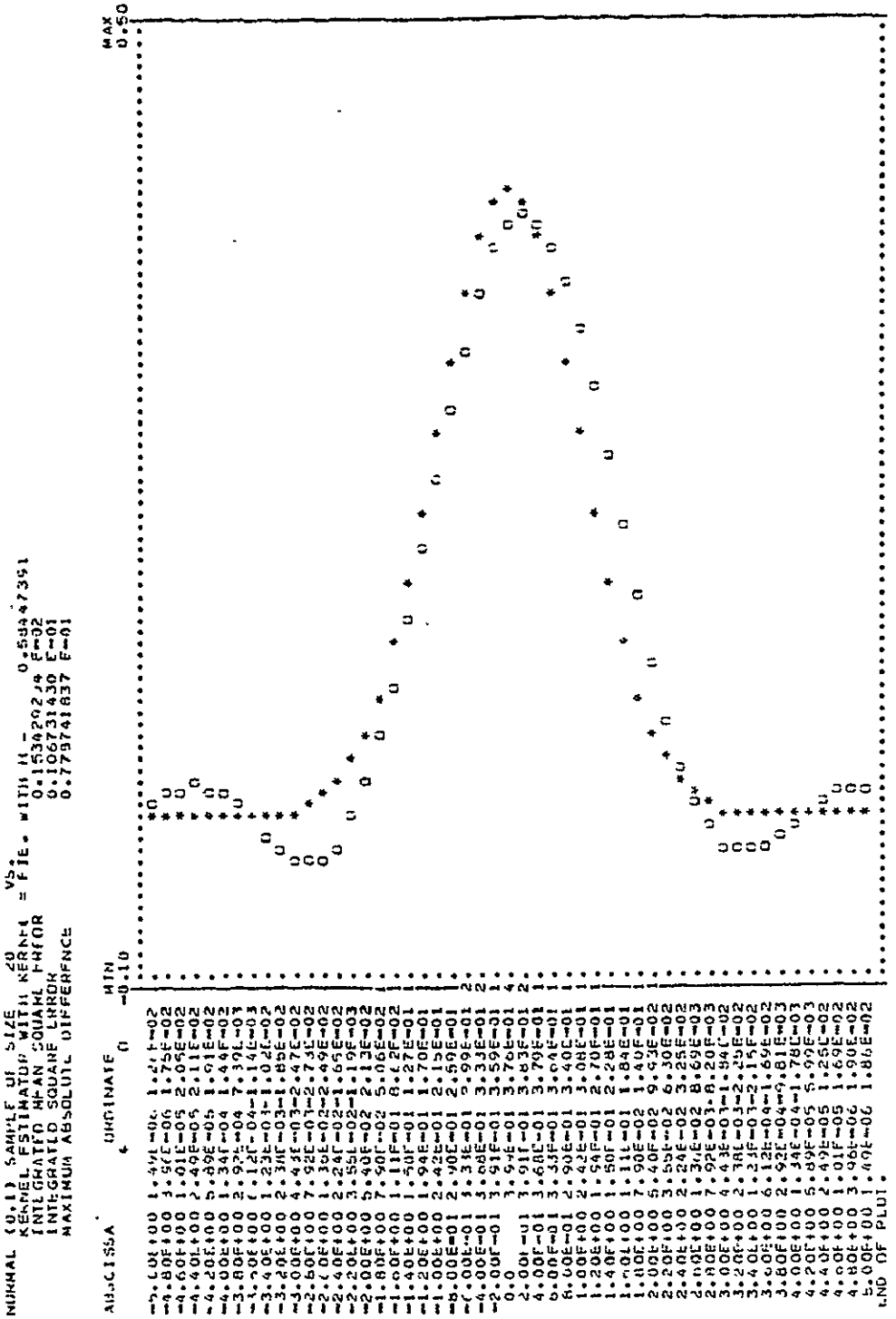


DIAGRAM 5.4.2c. N = 20 N(0,1) Quartic Kernel h(N) = 1.53

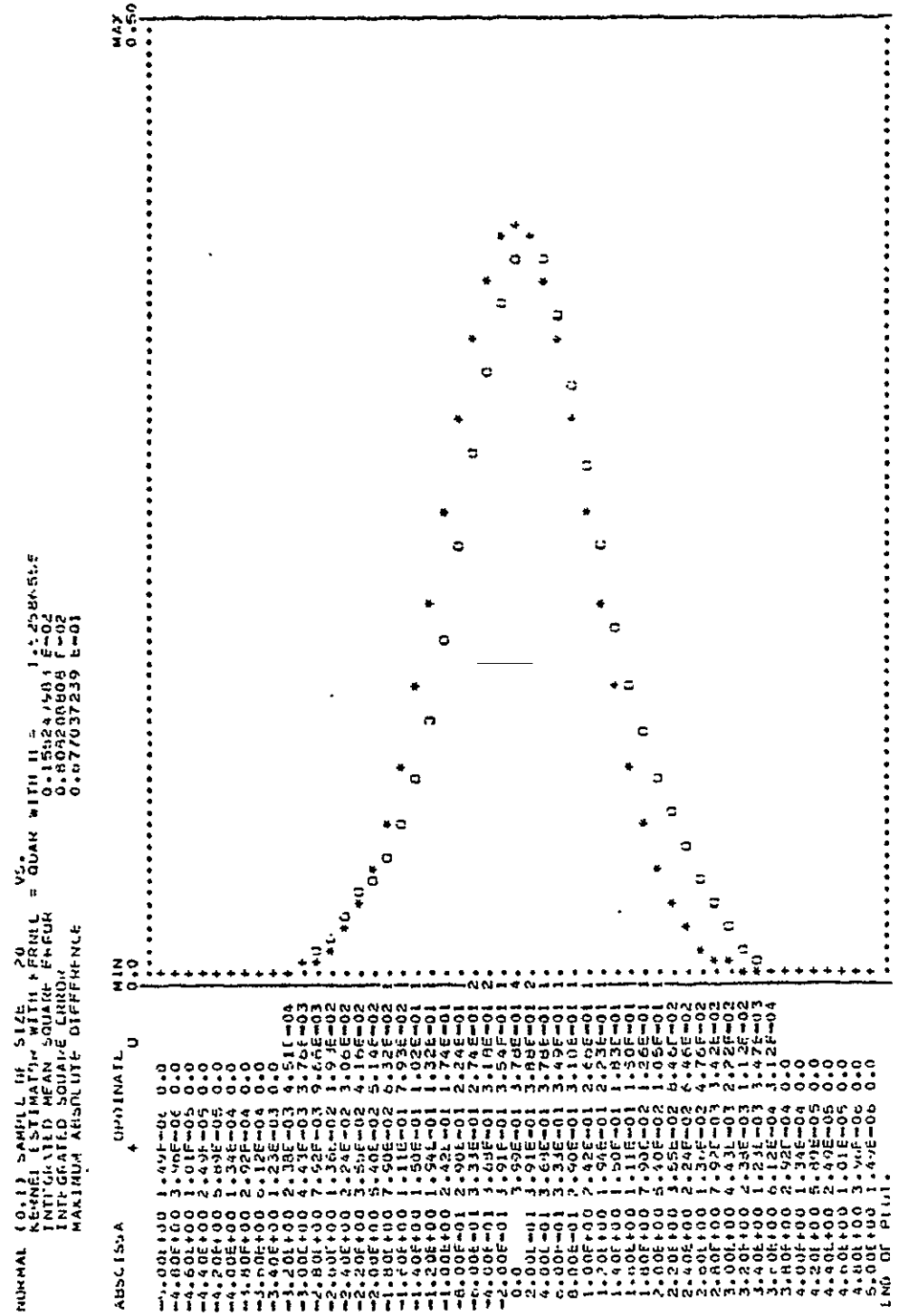


DIAGRAM 5.4.3a. N = 100 N(0,1) D.M.P.L.E. $\alpha = 10$

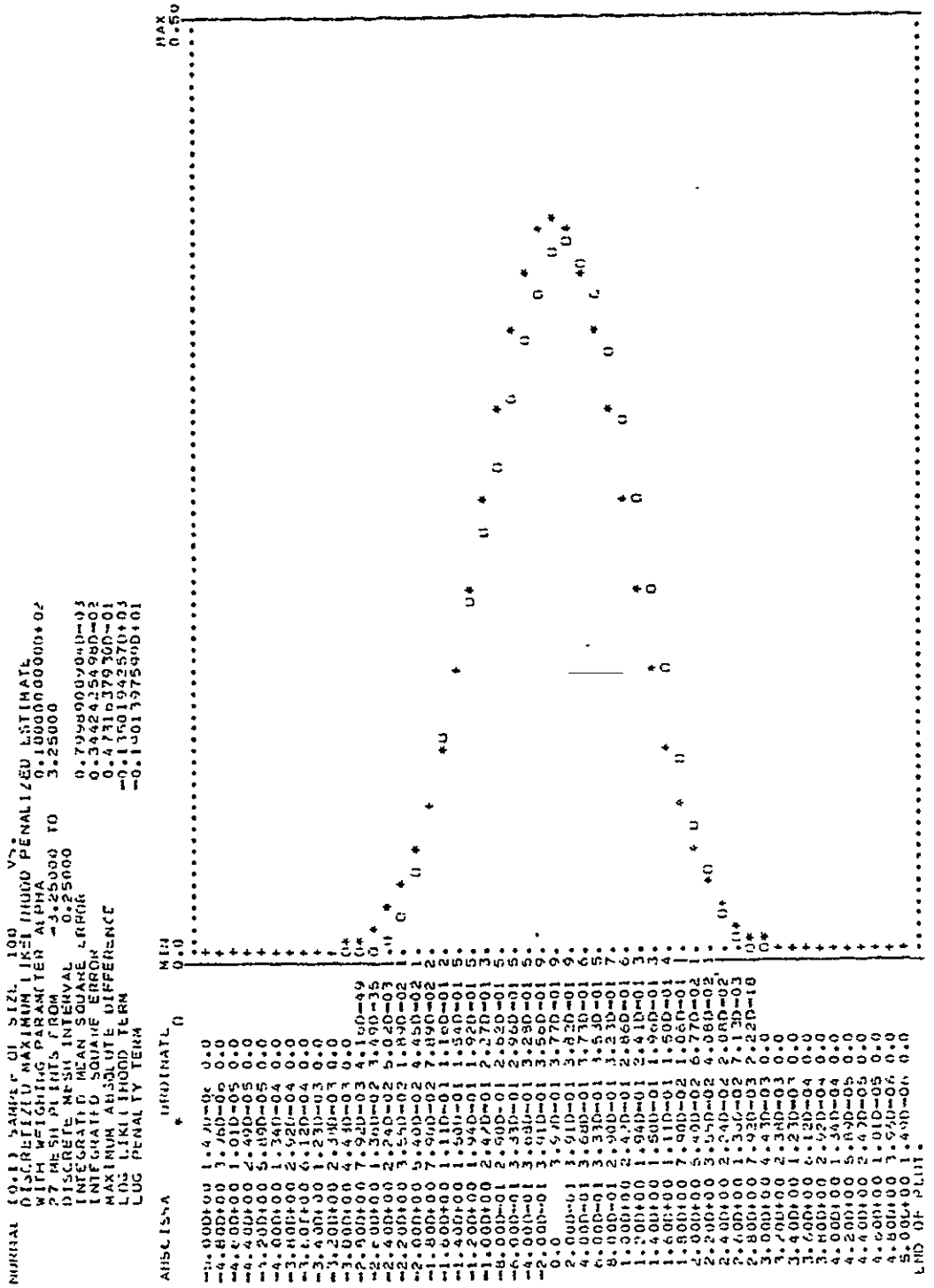


DIAGRAM 5.4.3b. N = 100 N(0,1) F.I.E. Kernel h(N) = 0.47

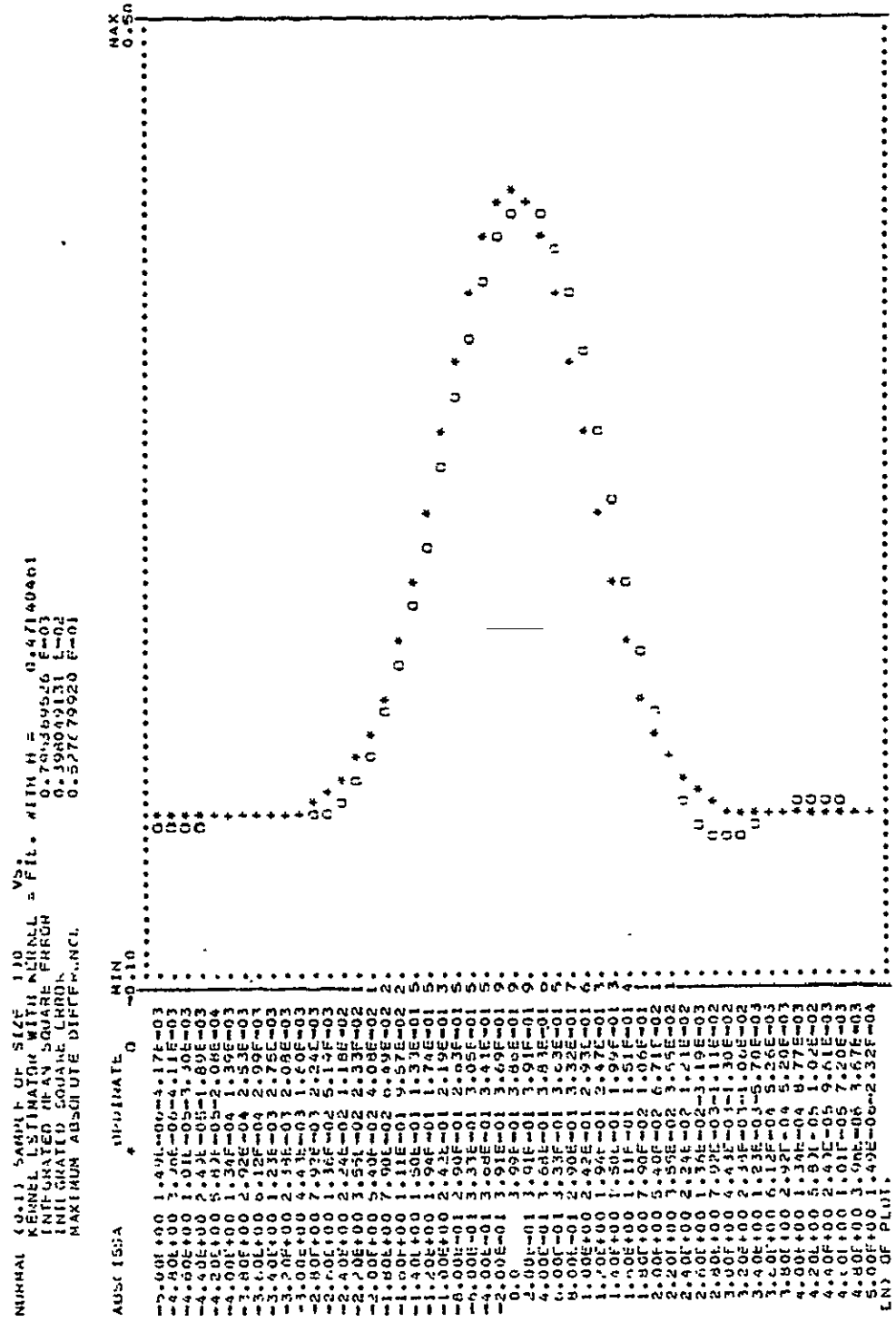


DIAGRAM 5.4.3c. N = 100 N(0,1) Quartic Kernel h(N) = 1.11

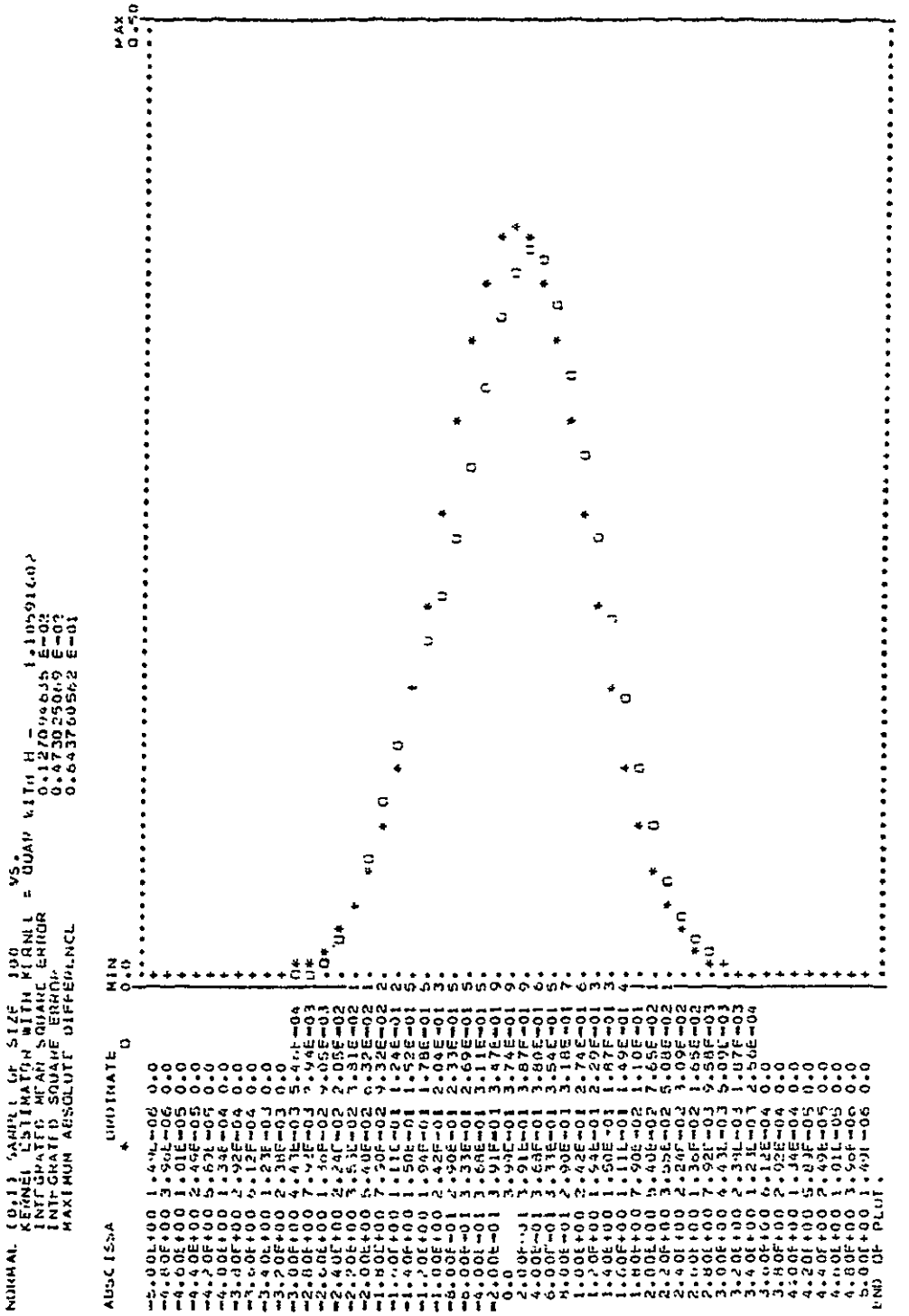


DIAGRAM 5.4.4a. N = 100 N(0,1) D.M.P.L.E. α = 10

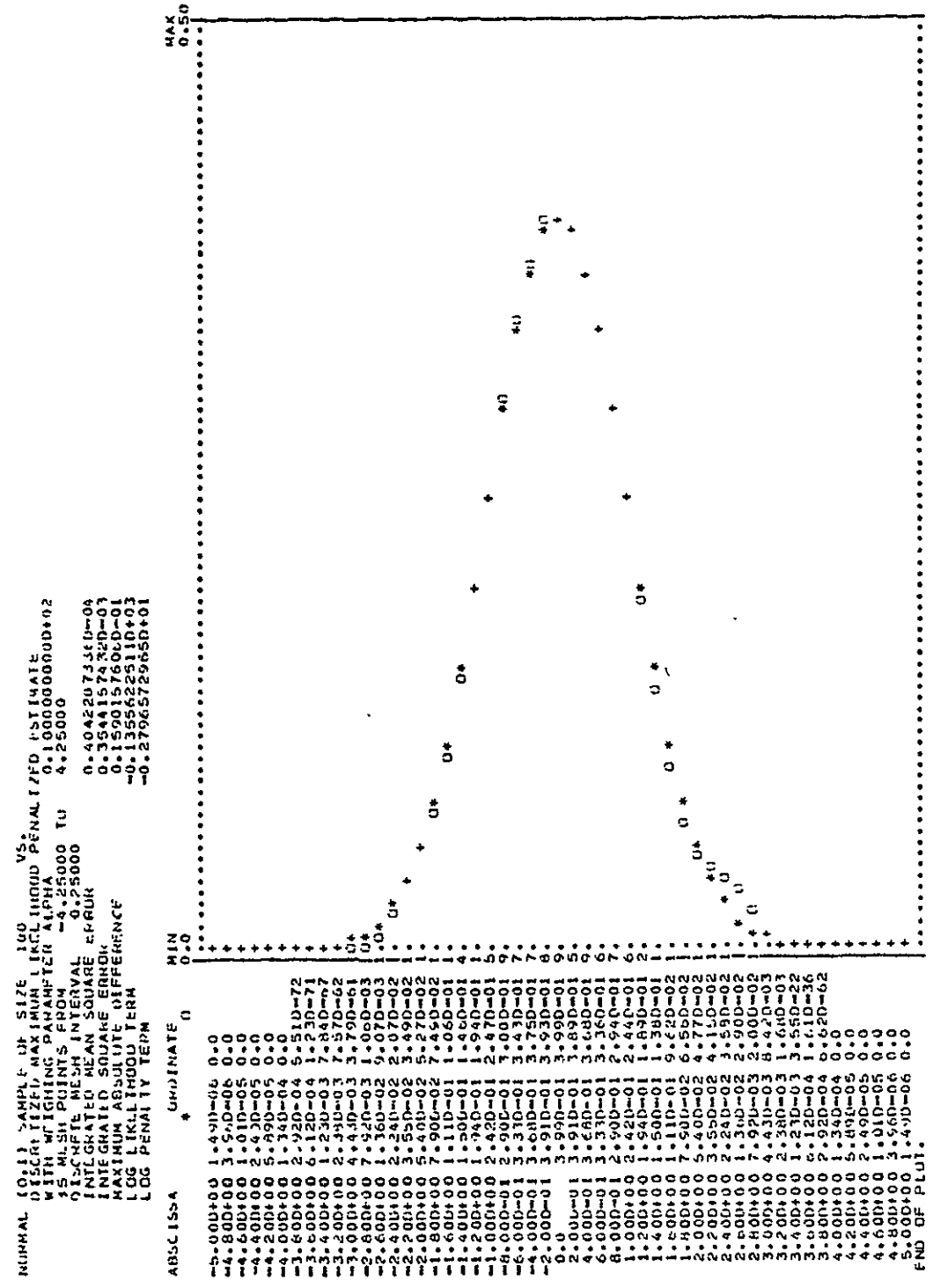


DIAGRAM 5.4.4b. N = 100 N(0,1) F.I.E. Kernel h(N) = 0.47

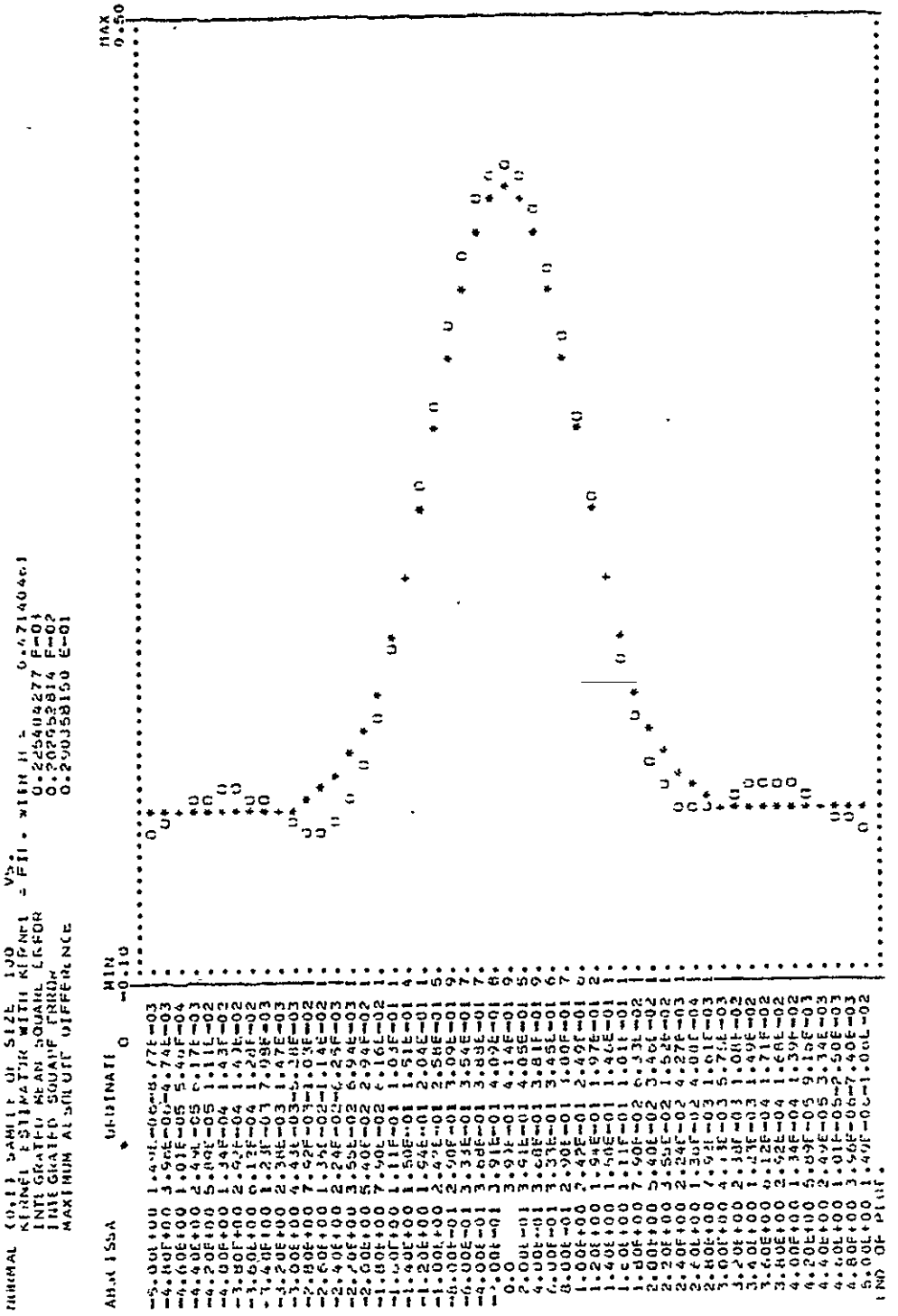


DIAGRAM 5.4.4c. N = 100 N(0,1) Quartic Kernel h(N) = 1.11

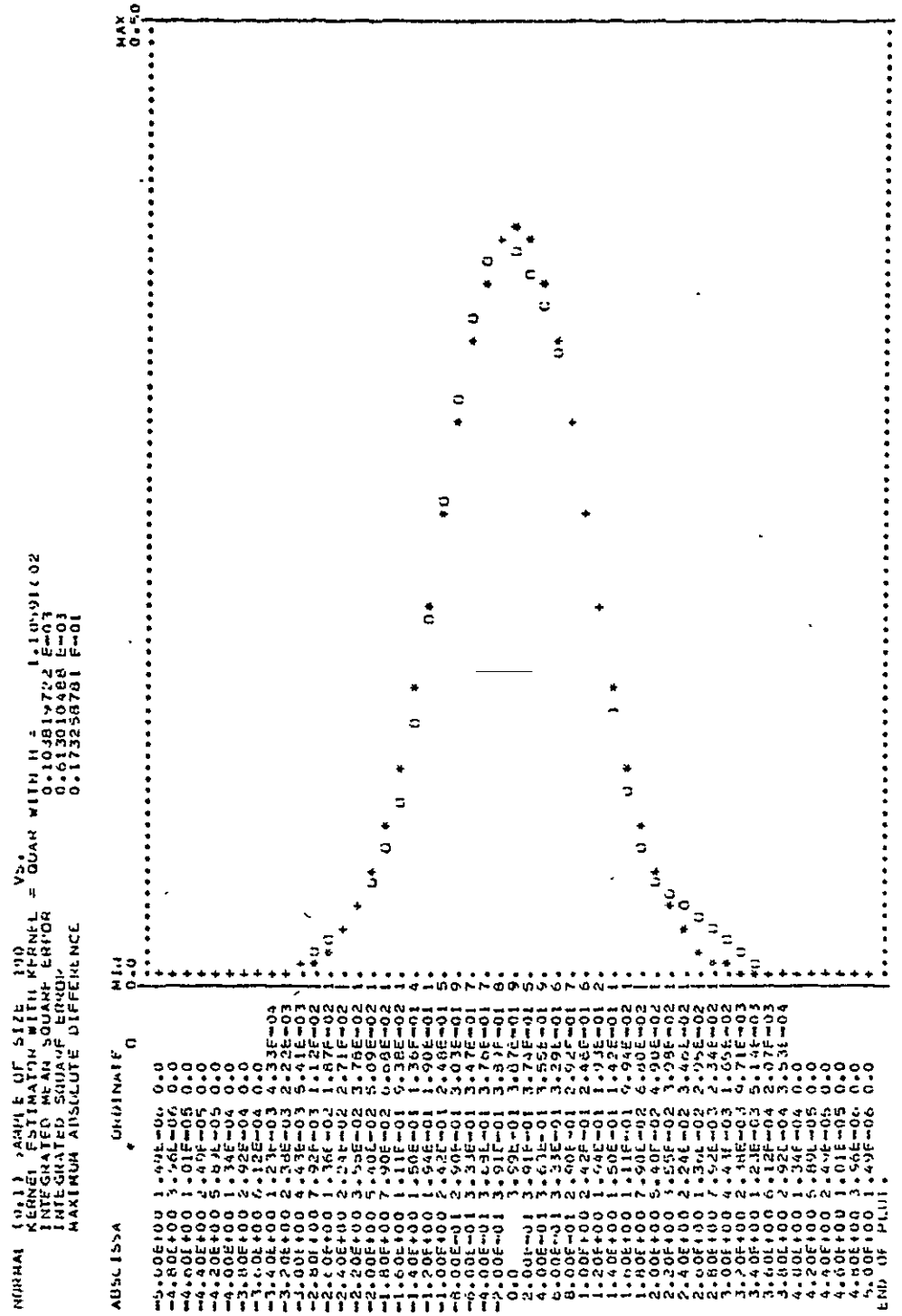


DIAGRAM 5.4.5a N = 25 Bimodal D.M.P.L.E. $\alpha = 10$

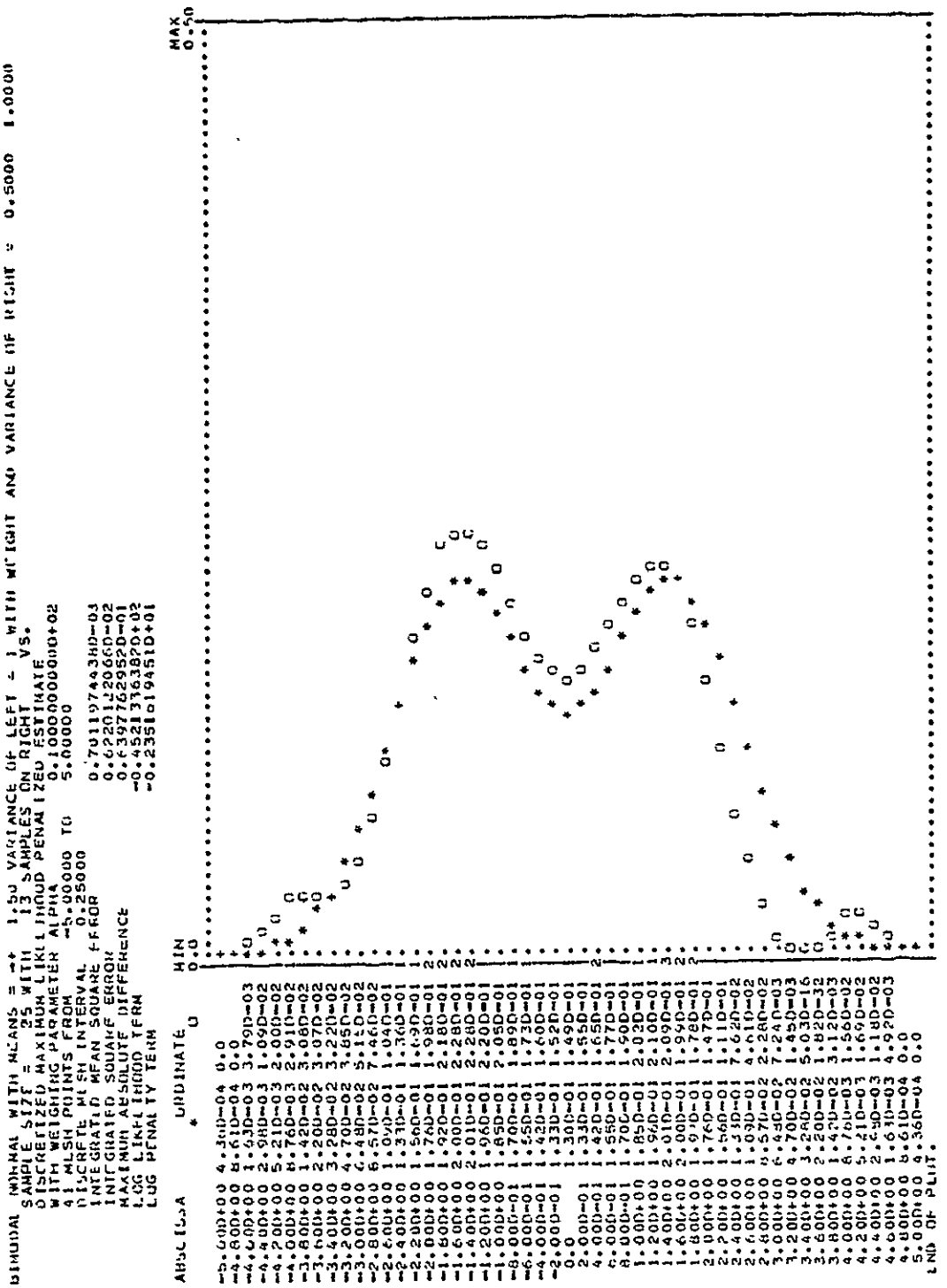


DIAGRAM 5.4.5b. N = 25 Bimodal F.I.E. Kernel h(N) = 0.56

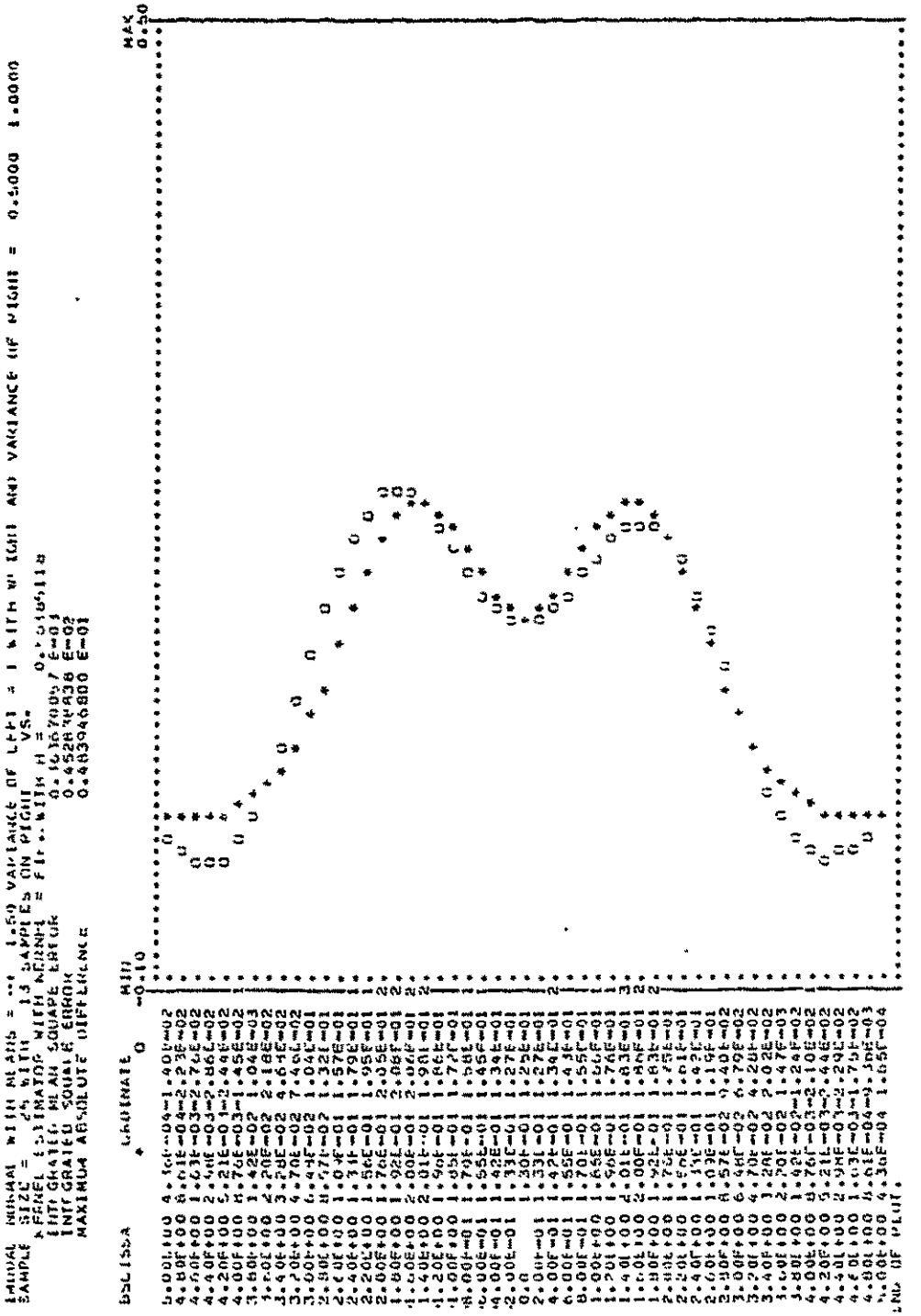


DIAGRAM 5.4.5c. N = 25 Bimodal Quartic Kernel h(N) = 1.72

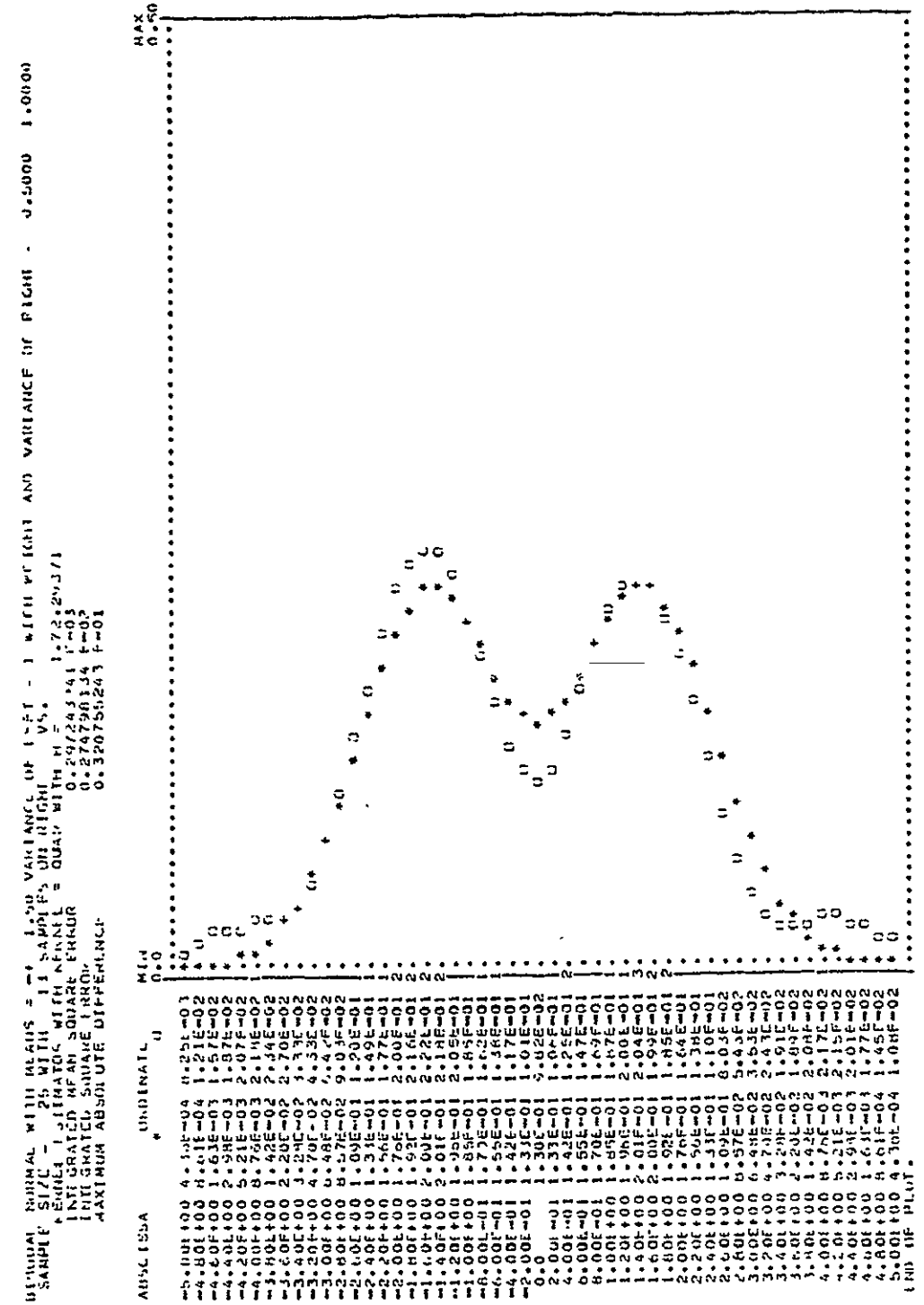
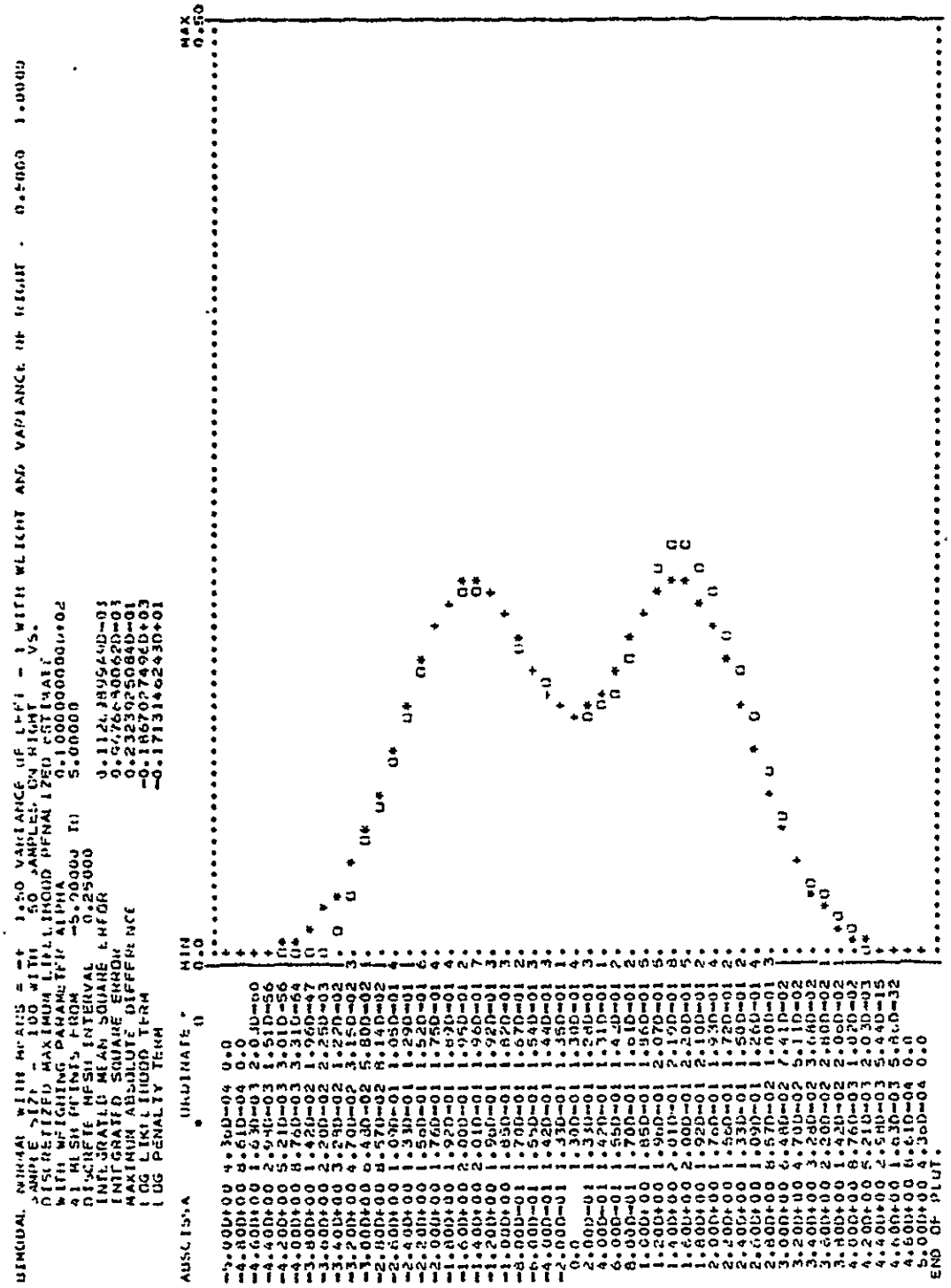


DIAGRAM 5.4.6a N = 100 Bimodal D.M.P.L.E. $\alpha = 10$



ORIGINAL PAGE IS
OF POOR QUALITY

DIAGRAM 5.4.6b. N = 100 Bimodal F.I.E. Kernel h(N) = 0.47

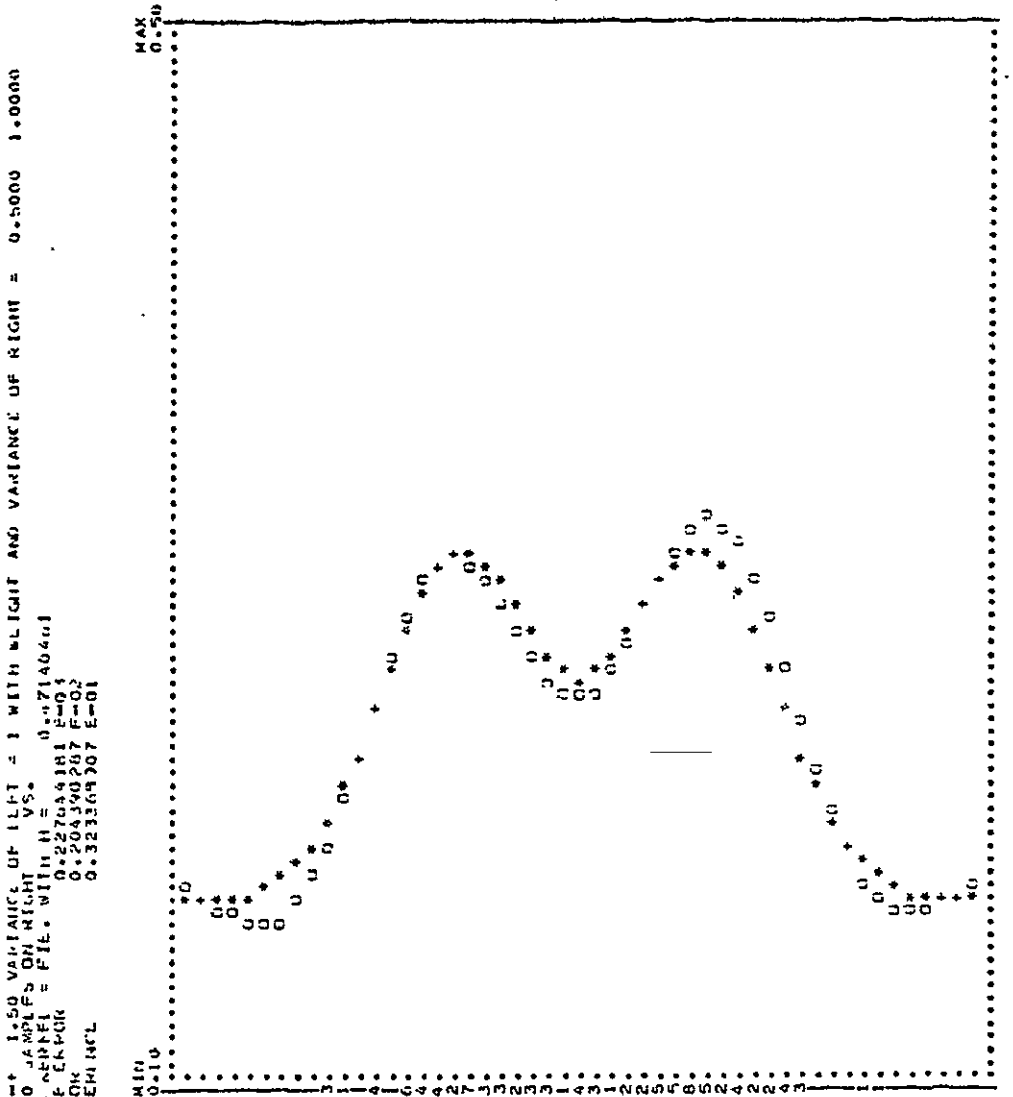


DIAGRAM 5.4.6c. N = 100 Bimodal Quartic Kernel h(N) = 1.31

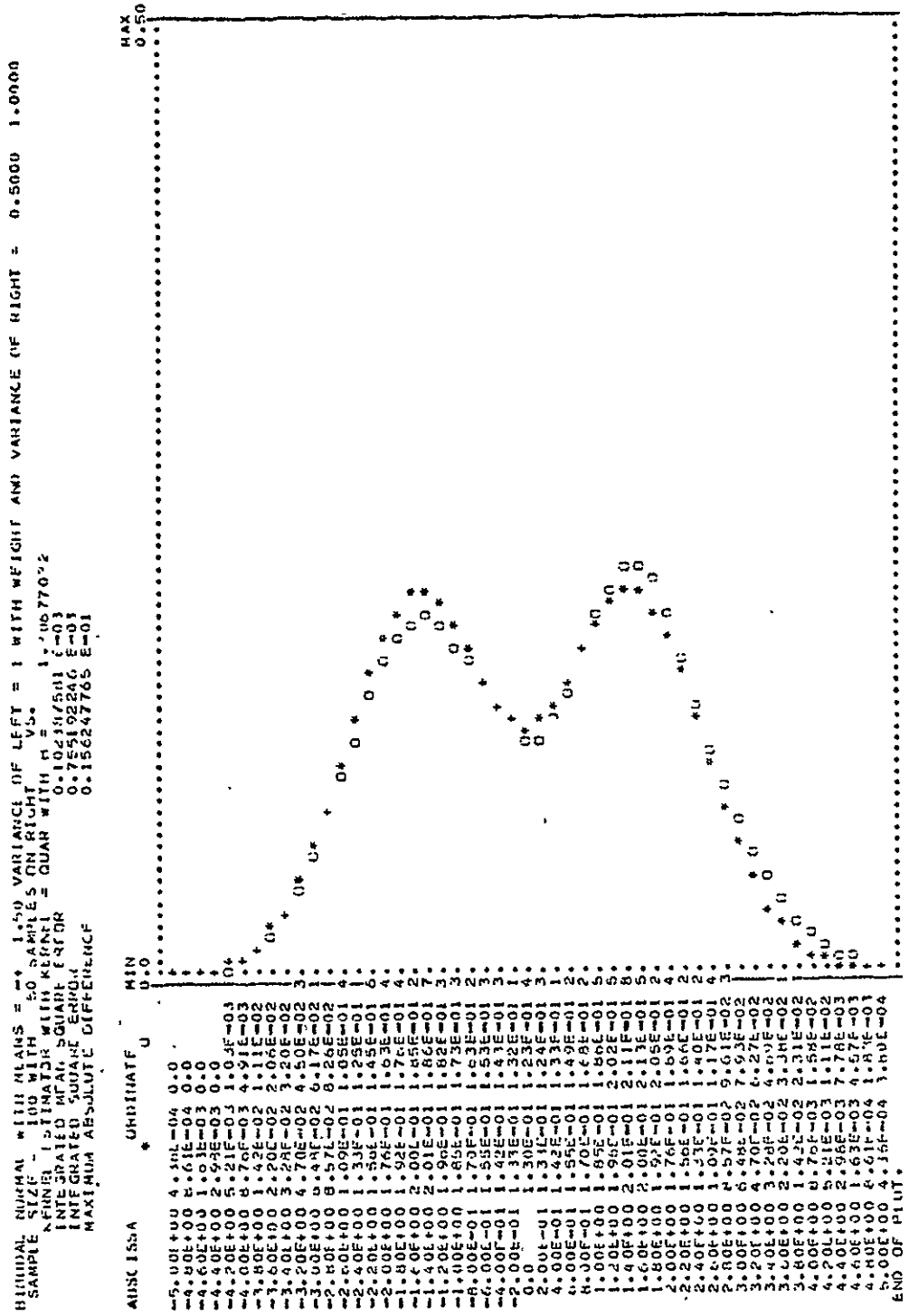


DIAGRAM 5.4.7a. $N = 400$ $F_{10,10}$ D.M.P.L.E. $\alpha = 0.5$

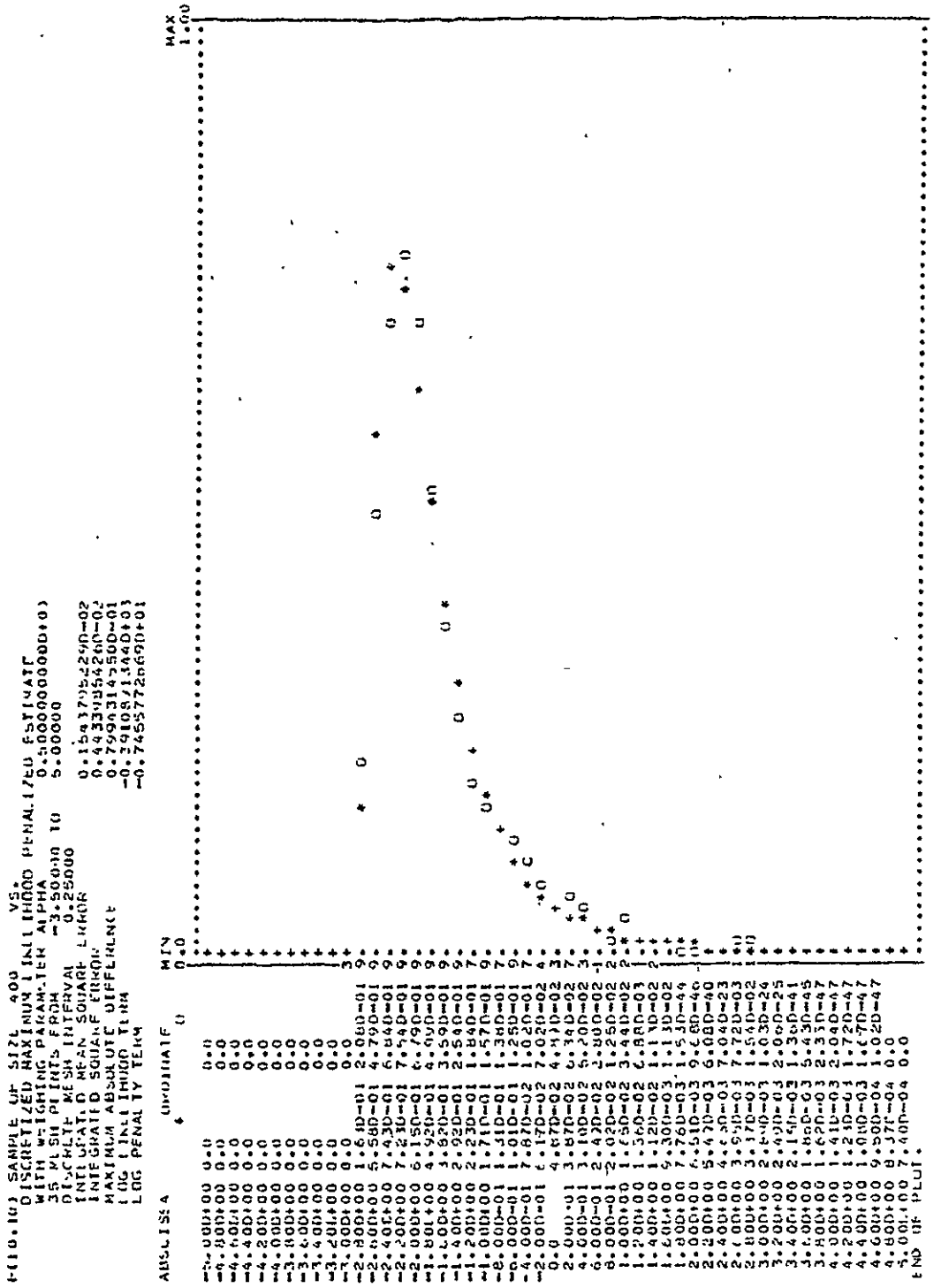


DIAGRAM 5.4.7b N = 400 $F_{10,10}$ F.I.E. Kernel h(N) = 0.41

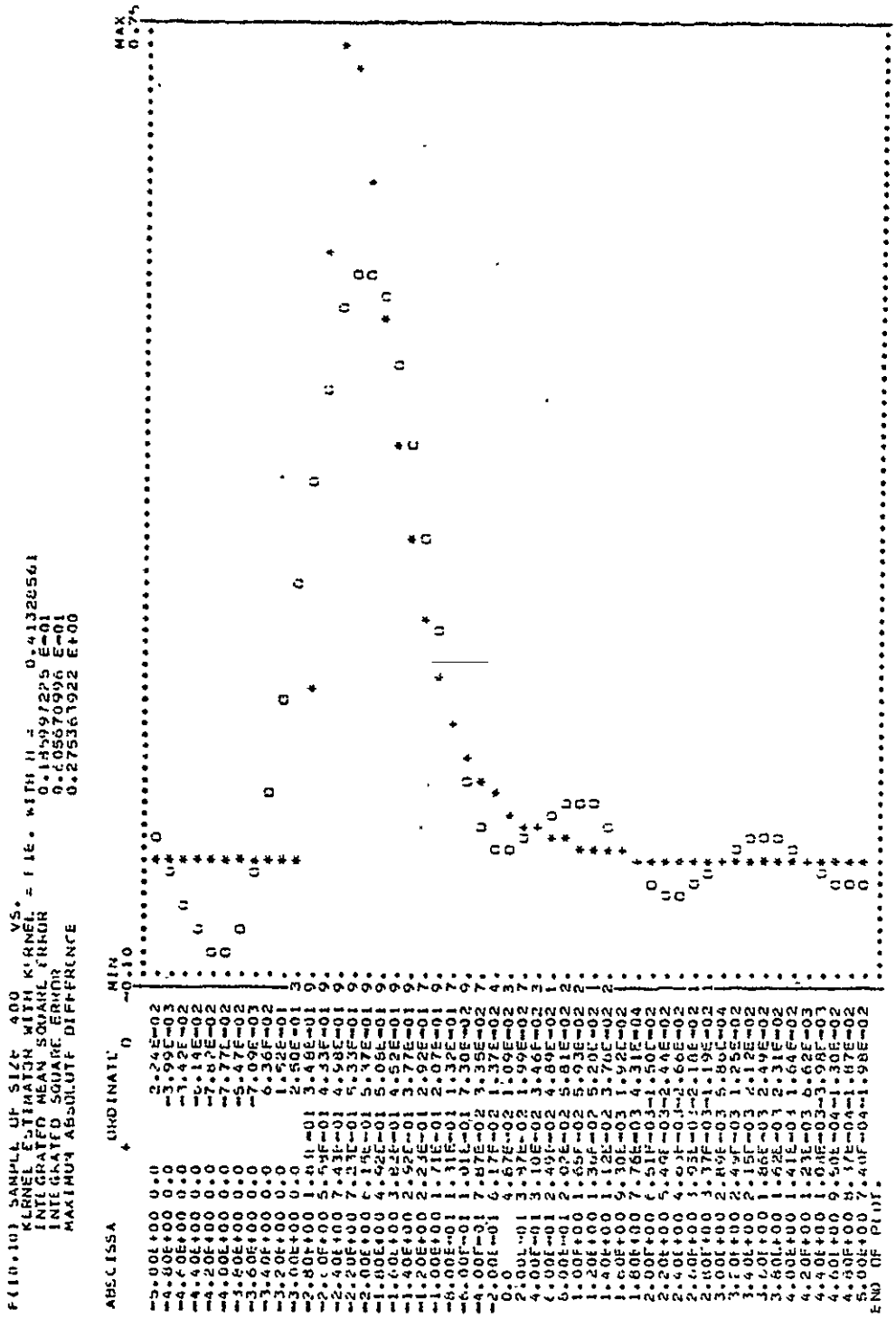


DIAGRAM 5.4.7b'. N = 400 $F_{10,10}$ F.I.E. Kernel h(N) = 0.25

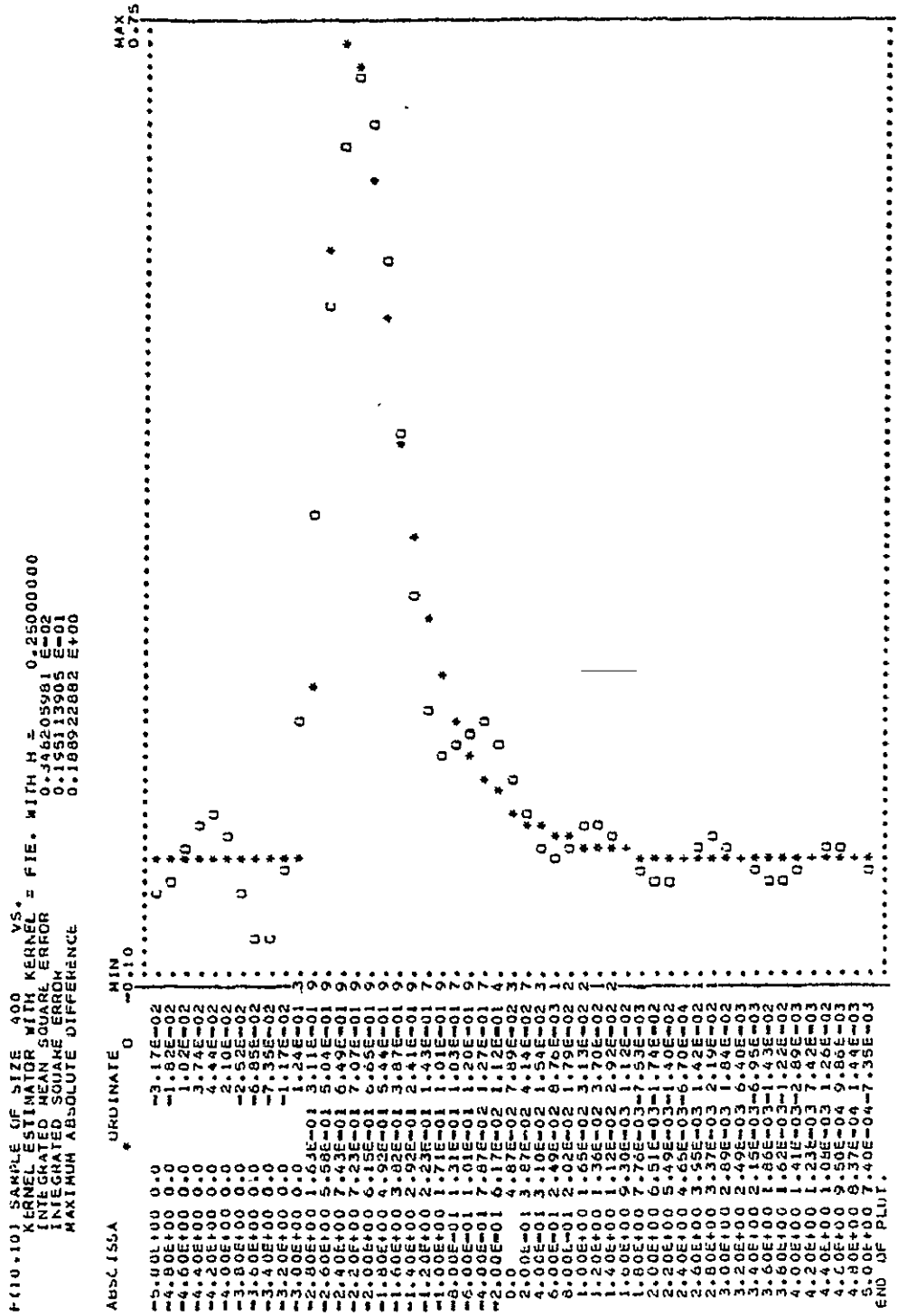


DIAGRAM 5.4.7b". N = 400 $F_{10,10}$ F.I.E. Kernel h(N) = 0.10

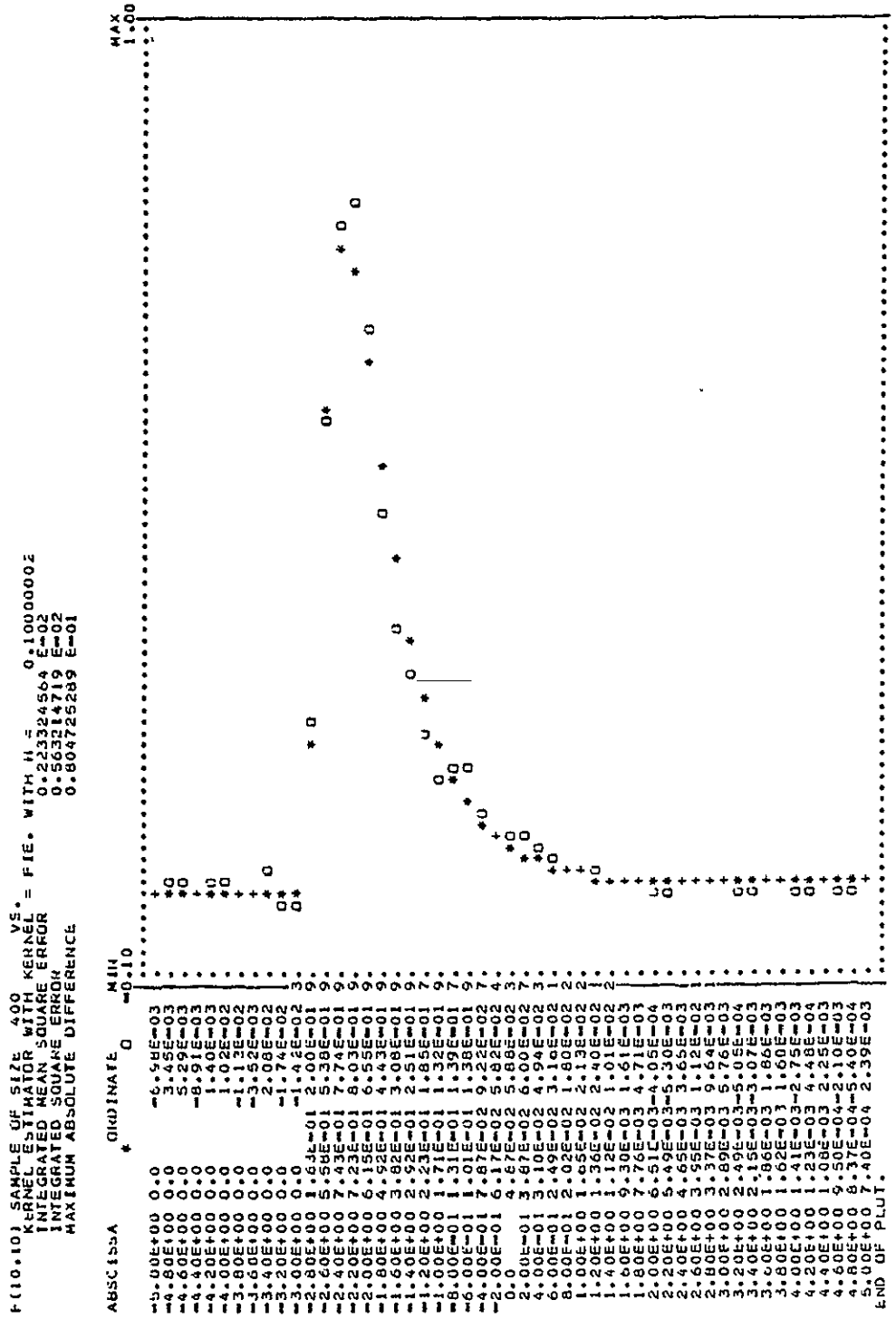


DIAGRAM 5.4.7b"'. N = 400 $F_{10,10}$ F.I.E. Kernel h(N) = 0.01

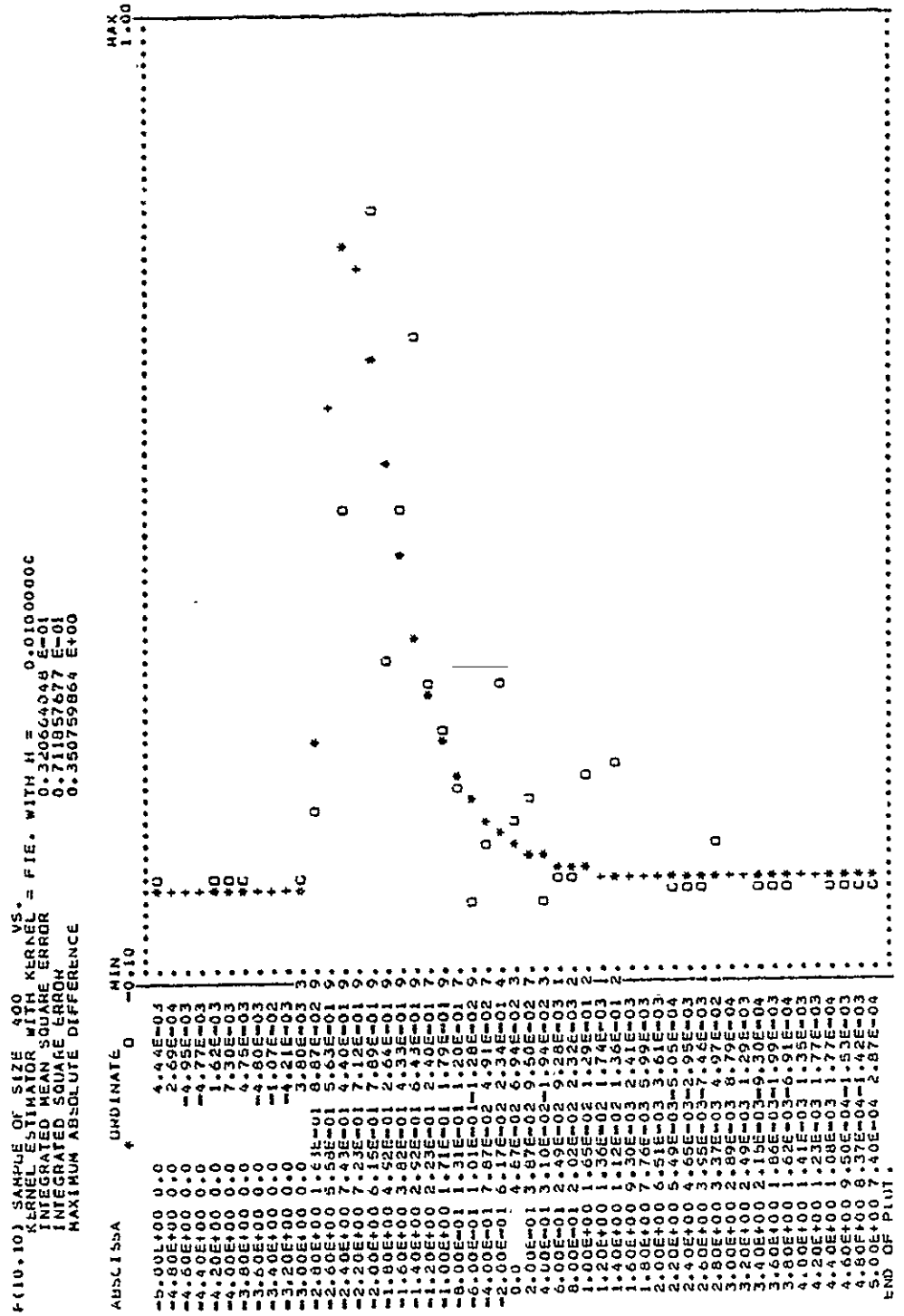
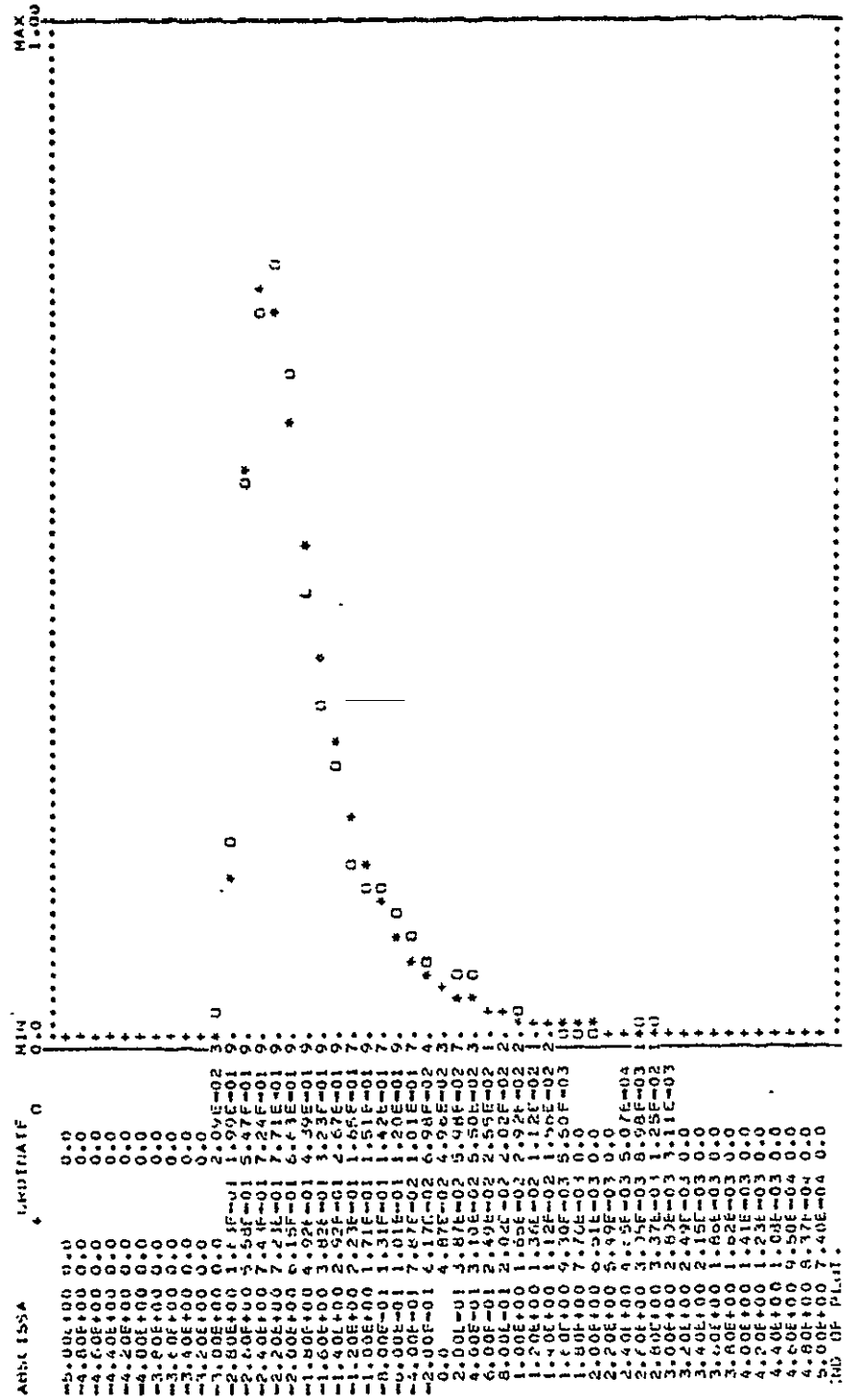


DIAGRAM 5.4.7c. N = 400 $F_{10,10}$ Quartic Kernel h(N) = 0.30



5.5 Monte Carlo Simulation Study

Random samples were generated from the densities (5.4.1) and the continuous piecewise linear discretized maximum penalized likelihood and kernel estimators calculated. Three measures of error were considered for an estimate $\hat{f}$ of f_o :

1. integrated mean square error

$$I.M.S.E. = \int_{-\infty}^{\infty} [\hat{f}(x) - f_o(x)]^2 f_o(x)dx \tag{5.5.1}$$

2. integrated square error

$$I.S.E. = \int_{-\infty}^{\infty} [\hat{f}(x) - f_o(x)]^2 dx \tag{5.5.2}$$

3. maximum absolute difference

$$DELMAX = \max_{x\in(-\infty,\infty)} |\hat{f}(x) - f_o(x)| \tag{5.5.3}$$

To evaluate (5.5.1)-(5.5.3) numerically, the values of $\hat{f}$ and f_o were calculated at the points -5.0, -4.9,...,0,...,4.9, 5.0 at a spacing of one-tenth. Simpson's rule was used to estimate (5.5.1) and (5.5.2). DELMAX was taken to be the maximum difference over the 101 points.

Twenty-five random samples were generated for each case discussed below for varying sample sizes. The quantities (5.5.1)-(5.5.3) were calculated for each sample. The mean and standard deviation were then calculated for quantities (5.5.1)-(5.5.3) using the 25 simulation results. To reduce computational time initial estimates in the maximum penalized likelihood algorithm were taken to be the true density values or one-hundredth, whichever was larger. Recall that an initial guess of zero in the numerical algorithm does not change in subsequent iterations. Typically about ten iterations were required to satisfy the convergence criterion (5.1.9). A rather coarse mesh interval h of 0.25 was chosen to reduce computational

times except in two instances where $h = 0.125$ was used. The mesh used is denoted by

$$\text{mesh} = (m, t_1, h) \tag{5.5.4}$$

where the estimate vanishes outside the interval (t_2, t_{m-1}). For the kernel estimators, "hopt" denotes the theoretically best choice given by (2.1.3) or (5.4.2). The Fourier integral estimate (F.I.E.) corresponds to the kernel (2.1.6). The quartic kernel is given in Table 2.5.1.

We remark that the quartic kernel exhibited smaller errors than did the Gaussian kernel. Using the Gaussian kernel increases computational time by a large factor with no apparent gain. The smoothness and finite support of the quartic kernel (and other spline kernels) are therefore attractive features.

The following computational times are typical for an IBM 370/155. For the discrete solution to generate and solve 25 samples from the four densities (5.4.1) with $N = 25, 100$ and 400 and with three values of α required 3439 seconds, about 3.82 seconds per sample for one value of α. For the Gaussian kernel estimate to generate and solve 25 samples from the four densities (5.4.1) with $N = 25$ and 100 for the optimal choice of $h(N)$ required 665 seconds, about 3.33 seconds per sample.

TABLE 5.5.1. Twenty-five N(0,1) samples each for N = 25, 100, 400, 800 (error means with standard deviations in parentheses)

N=25 Error	D.M.P.L.E.* α=10 mesh= (37,-4.5,.25)	F.I.E. Kernel hopt = .56	Quartic Kernel hopt = 1.46	Gaussian Kernel hopt = .56
I.M.S.E.	.0027 (.0019)	.0026 (.0021)	.0039 (.0031)	.0041 (.0032)
I.S.E.	.012 (.008)	.014 (.011)	.015 (.012)	.016 (.012)
DEIMAX	.077 (.031)	.079 (.027)	.091 (.039)	.095 (.039)

N=100 Error	D.M.P.L.E.* α=10 mesh= (37,-4.5,.25)	F.I.E. Kernel hopt = .47	Quartic Kernel hopt = 1.11	Gaussian Kernel hopt = .42
I.M.S.E.	.00079 (.00054)	.00085 (.00060)	.00122 (.00074)	.00129 (.00075)
I.S.E.	.0037 (.0021)	.0045 (.0026)	.0048 (.0027)	.0050 (.0027)
DEIMAX	.047 (.013)	.047 (.012)	.056 (.018)	.059 (.018)

N=400 Error	D.M.P.L.E.* α=10 mesh= (53,-3.25,.125)	F.I.E. Kernel hopt = .41	Quartic Kernel hopt = .84
I.M.S.E.	.00033 (.00018)	.00027 (.00020)	.00053 (.00022)
I.S.E.	.0014 (.0008)	.0013 (.0009)	.0020 (.0009)
DEIMAX	.031 (.008)	.025 (.009)	.039 (.010)

N=800	I.M.S.E.	I.S.E.	DEIMAX
D.M.P.L.E.* α=10 mesh= (53,-3.25,.125)	.00022 (.00013)	.0009 (.0005)	.026 (.006)

* 1,0,13,31 points were truncated for N = 25, 100, 400, 800 respectively. Three samples for N = 25 were calculated with the mesh = (53,-3.25,.125).

TABLE 5.5.2. Twenty-five Bimodal samples each for n = 25, 100, 400 (error means with standard deviations in parentheses)

N=25 Error	D.M.P.L.E.* α=10 mesh= (41,-5,.25)	Quartic Kernel hopt = 1.72	Gaussian Kernel hopt = .66
I.M.S.E.	.00159 (.00141)	.00120 (.00104)	.00128 (.00108)
I.S.E.	.012 (.010)	.008 (.007)	.009 (.007)
DELMAX	.071 (.030)	.061 (.022)	.063 (.023)

N=100 Error	D.M.P.L.E.* α=10 mesh= (41,-5,.25)	Quartic Kernel hopt = 1.31	Gaussian Kernel hopt = .50
I.M.S.E.	.00054 (.00032)	.00049 (.00031)	.00052 (.00031)
I.S.E.	.0040 (.0022)	.0034 (.0020)	.0036 (.0020)
DELMAX	.044 (.014)	.040 (.013)	.042 (.014)

N=400	I.M.S.E.	I.S.E.	DELMAX
D.M.P.L.E.* α=10 mesh= (41,-5,.25)	.00024 (.00012)	.0017 (.0007)	.030 (.007)

* 0,3,4 points were truncated for N = 25, 100, 400 respectively.

TABLE 5.5.3. Twenty-five t_5 samples each for N = 25, 100, 400 (error means with standard deviations in parentheses)

N=25 Error	D.M.P.L.E.* α=10 mesh = (41,-5,.25)	Quartic Kernel hopt = 1.07	Gaussian Kernel hopt = .41
I.M.S.E.	.00282 (.00148)	.00454 (.00229)	.00475 (.00233)
I.S.E.	.0147 (.0073)	.0203 (.0090)	.0210 (.0091)
DELMAX	.090 (.023)	.118 (.208)	.123 (.030)

N=100 Error	D.M.P.L.E.* α=10 mesh= (41,-5,.25)	Quartic Kernel hopt = .81	Gaussian Kernel hopt = .31
I.M.S.E.	.00084 (.00062)	.00150 (.00100)	.00157 (.00104)
I.S.E.	.0044 (.0027)	.0066 (.0038)	.0069 (.0039)
DELMAX	.048 (.017)	.068 (0.23)	.072 (.026)

N=400	I.M.S.E.	I.S.E.	DELMAX
D.M.P.L.E.* α=10 mesh= (41,-5,.25)	.00032 (.00020)	.0016 (.0008)	.032 (.009)

* 1, 17, 59 points truncated for N = 25, 100, 400 respectively.

TABLE 5.5.4 Twenty-five $F_{10,10}$ Samples Each for N = 25, 100, 400
(error means with standard deviations in parentheses)

N=25 Error	D.M.P.L.E.* α=.5 mesh = (35,-3.5,.25)	Quartic Kernel hopt = .52	Gaussian Kernel hopt = .20
I.M.S.E.	.0321 (.0270)	.0140 (.0104)	.0146 (.0105)
I.S.E.	.071 (.061)	.036 (.019)	.037 (.019)
DELMAX	.30 (.12)	.21 (.07)	.21 (.07)

N=100 Error	D.M.P.L.E.* α=.5 mesh = (35,-3.5,.25)	Quartic Kernel hopt = .39	Gaussian Kernel hopt = .15
I.M.S.E.	.0100 (.0071)	.0064 (.0049)	.0067 (.0051)
I.S.E.	.023 (.014)	.016 (.009)	.017 (.009)
DELMAX	.18 (.06)	.15 (.05)	.16 (.05)

N=400	I.M.S.E.	I.S.E.	DELMAX
D.M.P.L.E.* α = .5 mesh = (35,-3.5,.25)	.0029 (.0017)	.007 (.003)	.11 (.02)

* 2, 8, 21 points truncated for N = 25, 100, 400 respectively.

5.6 The Penalty Weighing Factor α

In this section we deal with two questions: first, whether $\alpha = \alpha(N)$ and second, how α is affected by scaling the random sample. The answer to the first question appears to be that α depends only on the underlying density and not on the sample size N. Good and Gaskins [1972, p. 188] give a heuristic proof that α is constant for the Normal density. In Table 5.6.1 we present the integrated mean square error for $\alpha = 5, 10, 20$ for the Normal, bimodal, and t_5 samples generated in the Monte Carlo study. For the $F_{10,10}$ samples $\alpha = \frac{1}{2}, 1, 2$ were used. We base our conclusions on these data. Perhaps a slight increase in α as N increases is indicated. However as is evident from Diagrams 5.3.1-5.3.6, dramatic changes in the estimates occur only for changes in magnitudes of α in powers of ten. This is due to the fact that the penalty term is competing against a logarithmic term that is less sensitive to small changes in α. In Table 5.6.2 we present a similar format for perturbing the optimal $h(N)$ for the kernel estimator with a Gaussian kernel.

A standard device is to transform the random sample $x_1,\ldots,x_N$ by

$$x_i' = \frac{x_i - a}{b} \qquad i = 1,N \tag{5.6.1}$$

for some choice of $a \in R$ and $b \in R_+$. Usually a is taken to be the sample mean and b the sample standard deviation. It is well known that this choice of a and b is not robust for densities with heavy tails. A more robust choice is

$$a = x_{(.5)} \tag{5.6.2}$$

$$b = \frac{2.16}{x_{(.86)} - x_{(.14)}}$$

where $x_{(p)}$ denotes the p^{th} sample quartile. The efficiency of (5.6.2)

TABLE 5.6.1. Average I.M.S.E. of the D.M.P.L.E for α Perburbed by a Factor of Two. Divide α by 10 for the $F_{10,10}$ Samples.

	I.M.S.E. for		
Sample	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$
N(0,1) N=25	.00242	.00267	.00427
N(0,1) N=100	.00093	.00079	.00089
N(0,1) N=400	.00037	.00033	.00035
N(0,1) N=800	.00028	.00022	.00019
Bimodal N=25	.00197	.00159	.00152
Bimodal N=100	.00070	.00054	.00171
Bimodal N=400	.00030	.00024	.00022
t_5 N=25	.00297	.00282	.00350
t_5 N=100	.00092	.00084	.00101
t_5 N=400	.00039	.00032	.00030
$F_{10,10}$ N=25	.03208	.03865	.05519
$F_{10,10}$ N=100	.00996	.01390	.02411
$F_{10,10}$ N=400	.00292	.00450	.00740

TABLE 5.6.2. Average I.M.S.E. of the Kernel Estimate for h(N) Perturbed by a Factor of Two (Gaussian kernel)

		I.M.S.E. for		
Sample	hopt	½ hopt	hopt	2 hopt
N(0,1) N=25	.556	.00804	.00411	.00843
N(0,1) N=100	.422	.00282	.00129	.00371
Bimodal N=25	.657	.00379	.00128	.00152
Bimodal N=100	.498	.00134	.00052	.00095
t_5 N=25	.406	.01067	.00475	.00416
t_5 N=100	.308	.00375	.00157	.00167
$F_{10,10}$ N=25	.198	.0334	.01456	.01999
$F_{10,10}$ N=100	.150	.01428	.00673	.00926

is less than that of (5.6.1) for a Normal sample, but not for a Cauchy or contaminated density.

If a transformation (5.6.1), a standard mesh $t'_1, t'_2, \ldots, t'_m$, and a reasonable choice for α' are used to solve for the continuous piecewise linear solution $p'(t')$, then the original problem has the solution

$$t_k = a + bt'_k$$
$$p(t) = \frac{1}{b} p'\left(\frac{t-a}{b}\right) \tag{5.6.3}$$

We ask whether $p(t)$ may be solved directly (given a and b) for some α.

Theorem 5.6.1. Suppose $t'_1, \ldots, t'_m$ is a fixed mesh in problem (5.1.4) with penalty weighing factor α'. Let the transformation constants a and b be given. Then the solution (5.6.3) for $p(t)$ may be solved directly by choosing $\alpha = b^{-5}\alpha'$ in problem (5.1.4) over the mesh $t_1, \ldots, t_m$.

Proof. The mesh spacing $h = bh'$. The integral constraint (5.1.5) is satisfied if

$$p(t_k) = \frac{1}{b} p'\left(\frac{t_k-a}{b}\right)$$

Problems (5.1.4) are

$$\text{maximize } \Sigma \log p(x_i) - \frac{\alpha}{h^3} \Sigma[\nabla^2 p(t_{k+1})]^2 \tag{5.6.4}$$

and

$$\text{maximize } \Sigma \log p'(x'_i) - \frac{\alpha'}{h'^3} \Sigma[\nabla^2 p'(t'_{k+1})]^2 . \tag{5.6.5}$$

Using (5.6.3), problem (5.6.4) becomes

$$\Sigma \log \frac{1}{b} p'(x'_i) - \frac{\alpha}{h^3} \Sigma[\frac{1}{b} \nabla^2 p'(t'_k)]^2$$

or

$$\Sigma \log p'(x_i') - N \log b - \frac{\alpha}{b^3 h'^3} \frac{1}{b^2} \Sigma [\nabla^2 p'(t_k')]^2 \quad .$$

Since $N \log b$ is constant, we have that the choice $\alpha' = b^{-5}\alpha$ renders problem (5.6.4) and (5.6.5) equivalent. This proves the theorem.

5.7 Extension to Higher Dimensions

For two-dimensional data the extension of the continuous piecewise linear function is the surface defined by triangles defined on a two-dimensional mesh. This problem is more difficult to solve. Another approach is the pseudo-independence algorithm of Bennett [1974]. After a linear transformation the problem of finding a p-dimensional density is reduced to that of finding p one-dimensional densities. Let x be a $p \times 1$ data vector. Let R be a $p \times p$ matrix and $\bar{x}$ a $p \times 1$ vector. Then the pseudo-independent estimate is

$$\hat{f}(x) = \prod_{i=1}^{p} \hat{f}_i(z_i) \qquad (5.7.1)$$

where

$$z = (z_1, z_2, \ldots, z_p)^t = R(x - \bar{x}) \qquad (5.7.2)$$

and $\hat{f}_i$ are p one-dimensional density estimates. Suppose $\bar{x}$ is the sample mean of the p-dimensional data and Σ is the positive definite sample covariance matrix. Let Λ denote the $p \times p$ diagonal matrix of the eigenvalues of Σ and E denote the corresponding $p \times p$ matrix of normalized eigenvectors. If we take

$$R = \Lambda^{-\frac{1}{2}} E^T \qquad (5.7.3)$$

then the transformed (5.7.1) data has mean zero and covariance matrix equal to the identity matrix. Only for Gaussian data does the product (5.7.1) have a theoretical justification. The pseudo-independence algorithm uses

(5.7.1) for arbitrarily distributed data.

For $p = 2$ the histogram and pseudo-independent discrete estimate with mesh interval 0.2 and $\alpha = 1.0$ are graphed for two data sets in Diagrams 5.7.1-5.7.4. The data are a measure of the intensity of light (reflected by the earth and recorded by satellite) in two spectral bands. The first band is from 0.40-0.44 μm and the second band is from 0.72-0.80 μm. The first data set is 225 pairs of measurements for light reflected from a soybean field. The second data set is 156 measurements on a corn field. For each data set a histogram is given to locate the random samples followed by the corresponding pseudo-independent discrete solution. Increasing values of the estimated two-dimensional density are denoted by the following ten symbols on a linear scale:

(smallest) 0 . , - / + ; * B $ (largest)

The parameters of the pseudo-independence algorithm are:

Data Set I

$$\bar{x} = \begin{pmatrix} 82.45 \\ 90.92 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 17.63 & 0 \\ 0 & 2.03 \end{pmatrix} \qquad E = \begin{pmatrix} .056 & .998 \\ .998 & -.056 \end{pmatrix}$$

Data Set II

$$\bar{x} = \begin{pmatrix} 85.48 \\ 103.17 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 26.28 & 0 \\ 0 & 5.79 \end{pmatrix} \qquad E = \begin{pmatrix} .323 & .946 \\ -.946 & .323 \end{pmatrix}$$

DIAGRAM 5.7.1. Histogram of 225 Soybean Data

x-range : 77.2 to 89.2 (by 1.2 ≡ 6 columns)

y-range : 77.9 to 104.5 (by 1.4 ≡ 3 rows)

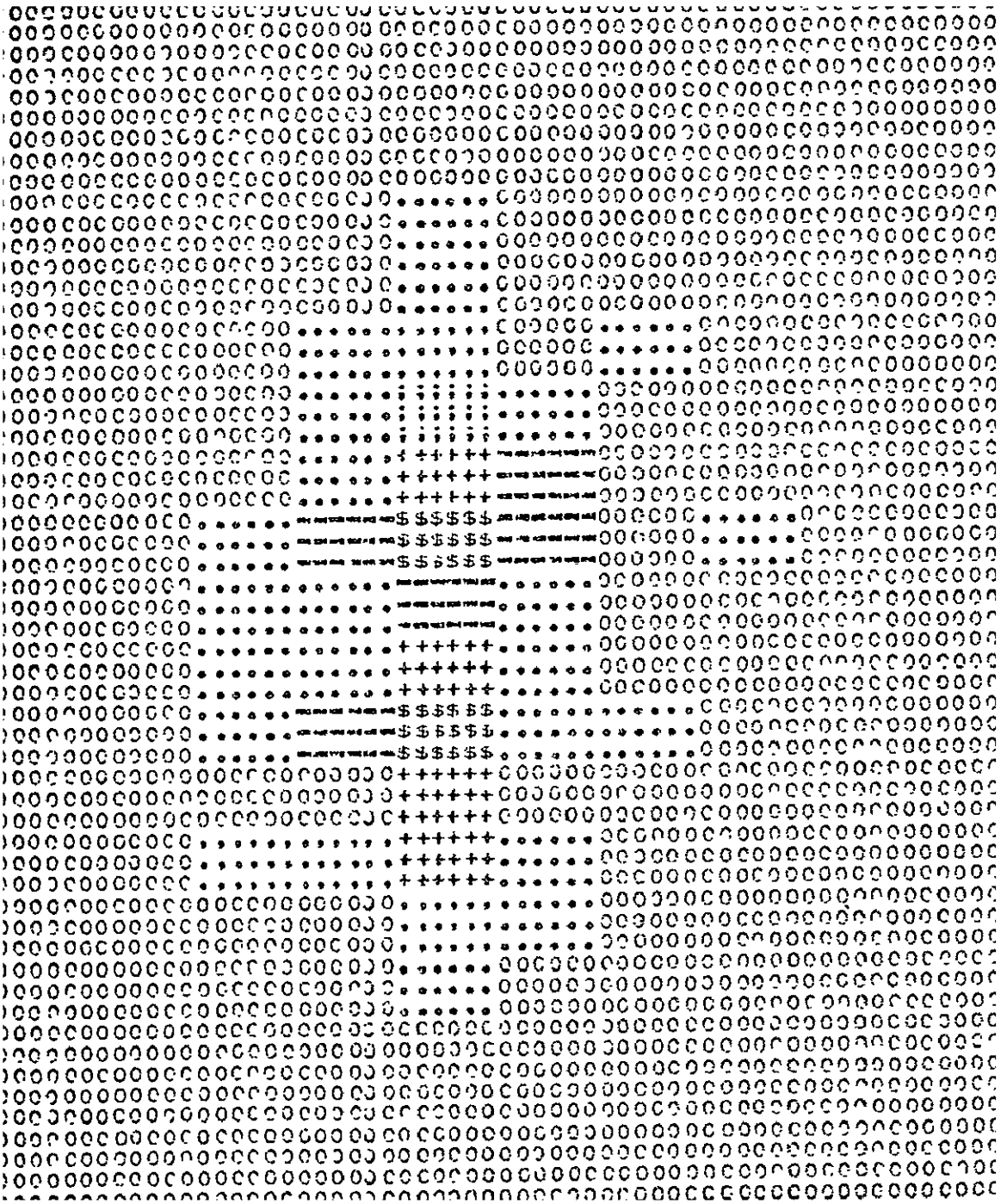


DIAGRAM 5.7.2. Pseudo-Independent Discrete Estimate of 225 Soybean Data $\alpha = 1$

x-range : 77.2 - 89.2 (by 0.2 ≡ 1 column)

y-range : 77.9 - 104.5 (by 7/15 ≡ 1 row)

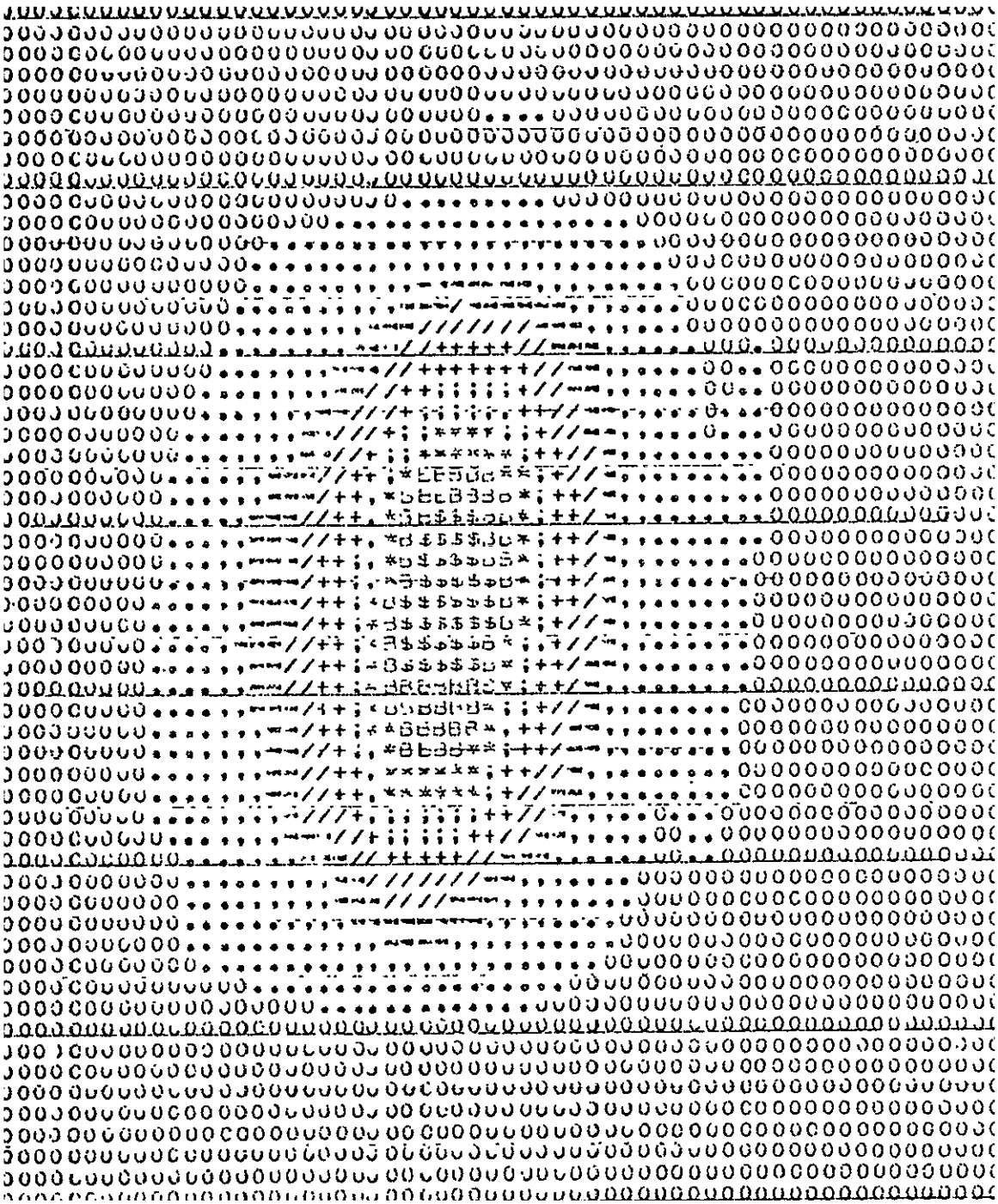


DIAGRAM 5.7.3. Histogram of 156 Corn Data

x-range : 76.75 - 94.75 (by 1.5 ≡ 5 columns)

y-range : 86.45 - 118.75 (by 1.7 ≡ 3 columns)

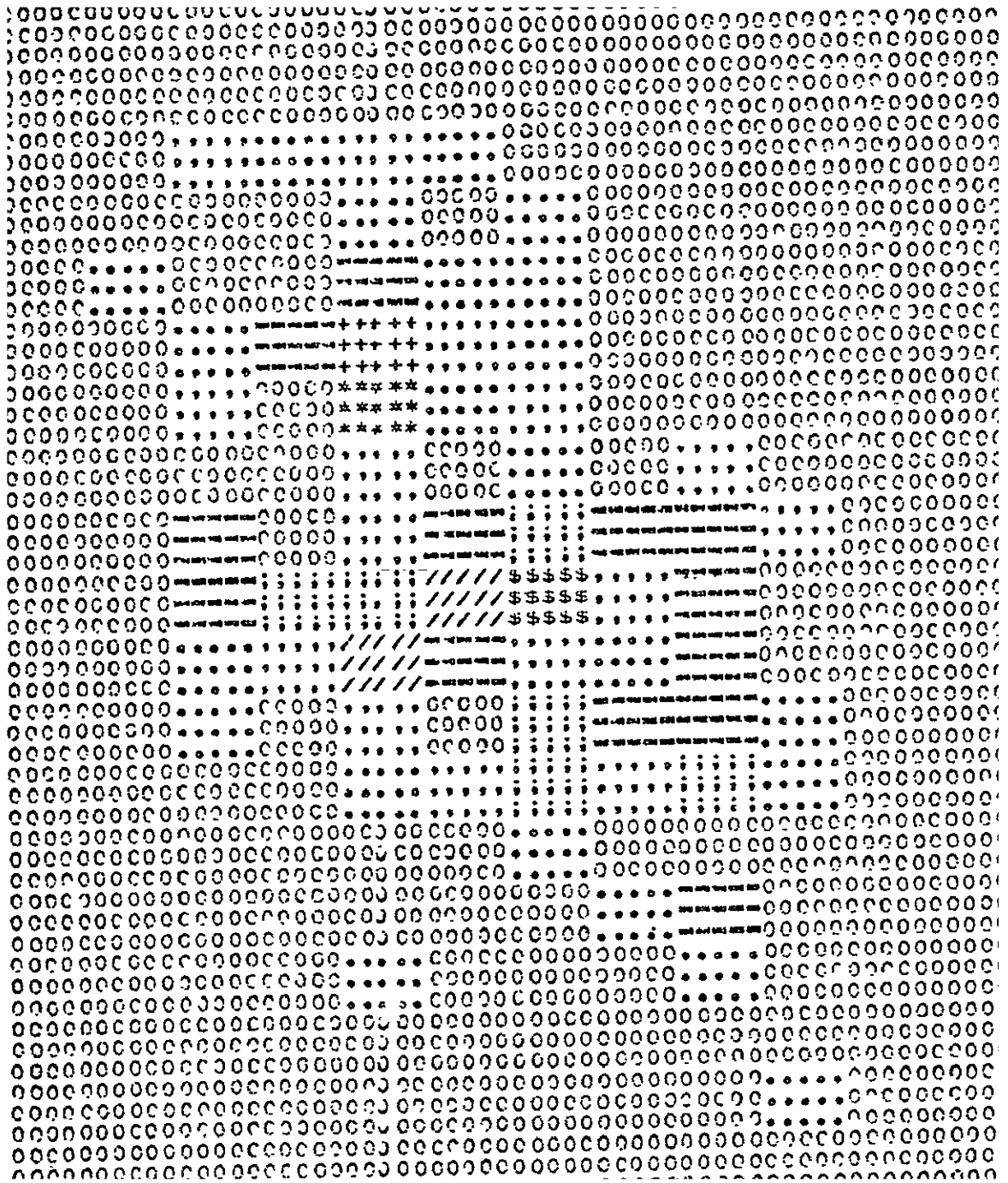
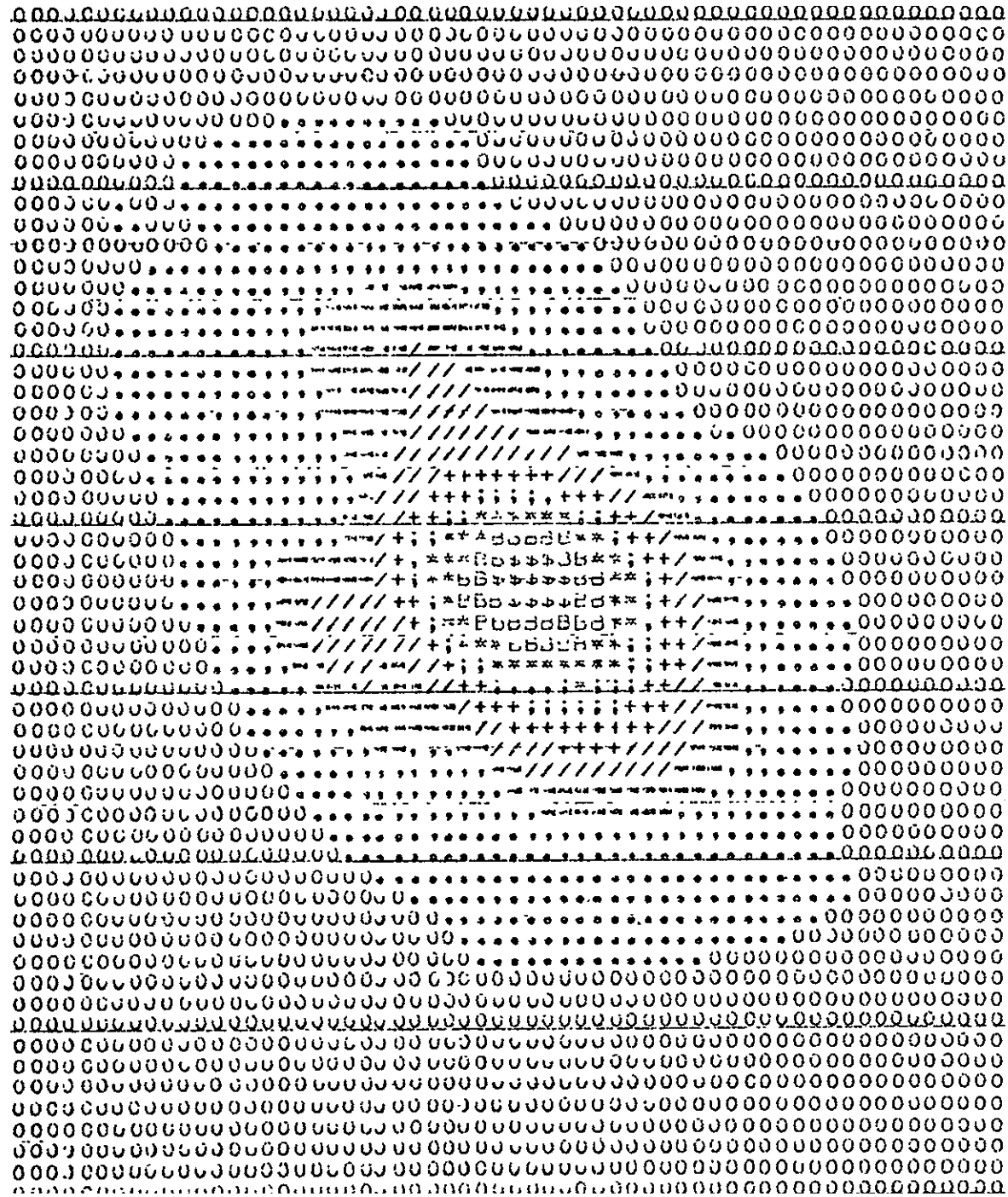


DIAGRAM 5.7.4. Pseudo-Independent Discrete Estimate of 156 Corn Data $\alpha = 1$

x-range : 76.75 - 94.75 (by 0.3 ≡ 1 column)

y-range : 86.45 - 118.75 (by 17/30 ≡ 1 row)



5.8 Conclusions

In this study, two nonparametric probability density estimation algorithms have been examined. The kernel estimators of Rosenblatt [1956] and Parzen [1962] are considered in Chapter II. The consistency properties of the kernel estimators are well known. The Fourier integral kernel of Davis [1975] is the most recent entry in this class of estimators; however, the resulting estimate is not nonnegative. Whittle's [1958] classical work attempts to find an optimal kernel to minimize the expected mean square error given prior information about the true density. A practical example that Whittle presents is corrected. The Whittle estimator is shown to be a Parzen kernel estimator when no prior information is available.

A difficulty with the kernel estimators is the choice of the kernel scaling parameter $h(N)$. An asymptotically optimal expression for $h(N)$ is known; however, a function of the true sampling density is required. In section 2.5 an interactive mode is described for choosing $h(N)$ using only the random sample and the investigator's prior feelings about the smoothness of the true sampling density. The interactive mode is extended to a proposed quasi-optimal algorithm for automatically picking $h(N)$ based solely on the data. In a Monte Carlo simulation study, the quasi-optimal estimate of $h(N)$ was obtained for randomly generated data sets. The integrated mean square error of the kernel estimate was calculated using the quasi-optimal and the theoretically optimal choices for $h(N)$. The efficiency of the quasi-optimal $h(N)$ was about 66%; however, the efficiency of the asymptotically optimal $h(N)$ scaled by a factor of two was less than 50%. Thus the quasi-optimal estimate performs well in light of the sensitivity of the kernel estimate to changes in $h(N)$. The obvious extension of the quasi-optimal algorithm to higher dimensions would be an

interesting exercise.

The second nonparametric probability density estimate is based on the maximum likelihood criterion. The histogram is shown to be the maximum likelihood estimator in the class of simple functions. In a more general class of functions, the maximum likelihood estimate may not exist; therefore, penalty function techniques are introduced in a natural way in a function space setting. In Chapter III, a theoretical basis is established for this class of estimators. Much of this material was motivated by a paper of de Montricher, Tapia, and Thompson [1975]. The maximum penalized likelihood estimate solves an infinite-dimensional problem and appears nontractable in general. Thus in Chapter IV a discrete version of the infinite-dimensional problem is introduced. The discretized maximum penalized likelihood estimator is shown to be a consistent in the mean square error. For a fixed sample the discrete solution approximates the infinite-dimensional solution as the mesh spacing approaches zero. Thus the discretized maximum penalized likelihood estimate is more robust in the choice of a mesh spacing than the histogram or the kernel estimate (with respect to the kernel scaling factor). Numerical studies have indicated that the D.M.P.L.E. does not change noticeably for h smaller than some positive threshold value. Consequently, we hypothesize the consistency requirement that the mesh spacing approach zero slowly as the sample size increases is an artifact of our proof. In other words, the mesh spacing may be picked arbitrarily small independently of the sample size. It should then be a direct result that the infinite-dimensional solution is also consistent. Open problems at this time include the rate of convergence of the discrete solution, the approximation properties of the discrete solution, and the proof of consistency for the original infinite-dimensional solution.

The numerical properties of the discrete solution are presented in Chapter V. Newton's method is employed to solve for the discrete estimate. An interactive mode is described for obtaining estimates given a random sample based on the investigator's prior feelings of the smoothness of the true sampling density. The robustness of the discrete estimator is demonstrated vis-a-vis the kernel estimator with respect to the choice of mesh, penalty weighing, and kernel scaling parameters. An extensive collection of graphs illustrates each of the ideas discussed. A Monte Carlo simulation study is summarized and a direct comparison made between the discrete and kernel estimators. The extension to density estimation in several dimensions is demonstrated by an example in two dimensions using data from NASA's Earth Resources project.

One important application for the discrete maximum penalized likelihood estimate is in the field of pattern recognition. The discrete maximum penalized likelihood estimator has advantages compared with the kernel estimator. The discrete solution does not involve the data for evaluation. In fact, the evaluation of the discrete estimate is as straightforward as a table lookup. On the other hand, the kernel estimator requires the data for evaluation, and the time required for evaluation increases with the sample size. Both the discrete and kernel estimates are superior to the Gaussian assumption for classification. The use of the Gaussian classifier requires as a preprocessing step the reduction of a class of training data into several subclasses of approximately Gaussian data.

The computational efficiency of the algorithm for calculating the discrete maximum penalized likelihood estimate can undoubtedly be improved. This efficiency is important when estimating multi-dimensional densities. The use of the pseudo-independence algorithm has appeared reasonably robust

against the multimodal possibilities encountered in remote sensing data. However, it is clear that this ad hoc projection of a p-dimensional density into p one-dimensional densities where the D.M.P.L.E. may be used will not be generally satisfactory. Thus it is clear that work needs to be carried out for generalizing the D.M.P.L.E. to the p-dimensional problem.

REFERENCES

Ash, R.B. (1972). Real Analysis and Probability. Academic Press, New York

Bennett, J.O. (1974). Estimation of multivariate probability density functions using B-splines. (Doctoral dissertation at Rice University, Houston, Texas).

Bennett, J.O., de Figueiredo, R.J.P., and Thompson, J.R. (1974). Classification by means of B-spline potential functions with applications to remote sensing. The Proceedings of the Sixth Southwestern Symposium on System Theory, FA3.

Boneva, L., Kendall, D.G., and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst. J.R. Statist. Soc. Ser. B., 33, 1 (including discussion).

Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Inst. Statist. Math., Tokyo. 18, 179.

Cencov, N.N. (1962). Evaluation of an unknown distribution density from observations. Soviet Math., 3, 1559.

Davis, K.B. (1975). Mean square error properties of density estimates. Ann. Statist., 3, 1025.

de Montricher, G.M. (1973). Nonparametric Bayesian estimation of probability densities by function space techniques. (Doctoral dissertation at Rice University, Houston, Texas).

de Montricher, G.F., Tapia, R.A., and Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Ann. Statist., 3, 1329.

Epanechnikov, V.A. (1969). Nonparametric estimates of a multivariate probability density. Theor. Prob. Appl., 14, 153.

Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. Phil. Trans., A, 222, 309.

Fisher, R.A. (1950). Contributions to Mathematical Statistics, Wiley, New York.

Fix, E. and Hodges, J.L., Jr. (1951). Discriminatory analysis, nonparametric discrimination, consistency properties. USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 4.

Good, I.J. and Gaskins, R.A. (1972). Global nonparametric estimation of probability densities. Virginia J. Sci., 23, 171.

Huzurbazar, V.S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. Ann. Eugen., 14, 185.

Kazakos, D. (1975). Optimal choice of the kernel function for the Parzen kernel-type density estimators. Submitted for publication.

Kendall, M.G. and Stuart, A. (1955). The Advanced Theory of Statistics, Vol. 1, Distribution Theory. Buttler and Tanner, London.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics, Vol. 2, Inference and Relationship. Griffin, London.

Noble, B. (1969). Applied Linear Algebra. Prentice-Hall, Englewood Cliffs, New Jersey.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., 33, 1065.

Parzen, E. (1967). Time Series Analysis Papers. Holden-Day, San Francisco.

Pearson, K. (1948). Karl Pearson's Early Statistical Papers. Cambridge Univ. Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27, 832.

Rosenblatt, M. (1971). Curve estimates. Ann. Math. Statist., 42, 1815.

Roussas, G.G. (1973). A First Course in Mathematical Statistics. Addison-Wesley, Reading, Massachusetts.

Schoenberg, I.J. (1973). Splines and histograms. Spline Functions and Approximation Theory, ed. A. Meir and A. Sharma, Birkhäuser Verlag Basel, Stuttgart.

Schultz, M.H. (1973). Spline Analysis. Prentice-Hall, Englewood Cliffs, New Jersey.

Stein, E.M. and Weiss, G. (1971). Introduction to Fourier Analysis on Euclidean Spaces. Princeton Univ. Press.

"Student" (1908). The probable error of a mean. Biometrika, 6, 1.

Tapia, R.A. (1974). A stable approach to Newton's method for general mathematical programming problems in R^n. J. Opt. Th. and Appl., 14, 453.

Wahba, G. (1971). A polynomial algorithm for density estimation. Ann. Math. Statist., 42, 1870.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist., 20, 595.

Watson, G.S. (1969). Density estimation by orthogonal series. Ann. Math. Statist., 40, 1496.

Watson, G.S. and Leadbetter, M.R. (1963). On the estimation of the probability density, I. Ann. Math. Statist., 34, 480.

Wegman, E.J. (1969). A note on estimating a unimodal density. Ann. Math. Statist., 40, 1661.

Wegman, E.J. (1970). Maximum likelihood estimation of a unimodal density function. Ann. Math. Statist., 41, 457.

Wegman, E.J. (1972). Nonparametric probability density estimation: I. a summary of available methods. Technometrics, 14, 533.

Wegman, E.J. (1976). Maximum likelihood estimation of a probability density function. To appear in Sankya.

Whittle, P. (1958). On the smoothing of probability density functions. J.R. Statist. Soc. (B) 20, 334.